ZCAS UNIVERSITY

2025

Design and Development of an Intelligent Framework for Human Resource Case Document Processing: Integrating Image Processing, OCR, NLP, Sentiment Analysis and Artificial Intelligence

WEDEX CHITALU ILUNGA

STUDENT NUMBER: G12114

A Final Year Research Project submitted in partial fulfilment of the requirements for the degree of

Master of Science in Computer Science

DECLARATION

Name: Wedex Chitalu Ilunga

Student Number: G12114

I hereby declare that this final year research project is the result of my own work, except for quotations and summaries which have been duly acknowledged.

Plagiarism check: %

Signature: Date:

Supervisor Name: Prof. Aaron Zimba

Supervisor Signature:

Date:

Design and Development of an Intelligent Framework for Human Resource Case Document Processing: Integrating Image Processing, OCR, NLP, Sentiment Analysis and Artificial Intelligence

ABSTRACT

This project is focused on the development and demonstration of an automated Human Resource (HR) case processing system designed to facilitate the storage and analysis of data to support decision making processes by extracting meaningful data from scanned HR case documents. Specifically, this project is intended for implementation in governmental or otherwise bureaucratic HR processes. Governmental context provides suitable conditions for the utilization of Optical Character Recognition (OCR) for the extraction of text and subsequent derivation of key aspects due to these documents maintaining a standard structured form with rare, minor deviations. Traditional manual, paper-based methods of human resource case document processing are commonplace within many organisations. Outdated systems such as these hinder efficiency and are prone to suffering the challenges of document deterioration, misplacement, and inaccuracies, leading to delays in decision-making, duplication of work, general errors in case resolution, and reduction in operational efficiency. To address the challenges posed by manual, paper-based document processing systems, an intelligent framework is proposed incorporating image pre-processing techniques for high accuracy scanning using computer vision for ROI detection and Tesseract OCR Engine for text identification and extraction. In addition, Sentiment Analysis is implemented for text processing based on keywords identified through RegEx to determine cases and extract data for structured storage in an SQL, ensuring referential posterity and non-biased decision making. An OCR Accuracy of 95% was successfully achieved—statistically, well-within standard benchmarks of OCR accuracy for the scanning of printed text. Further to this, the framework incorporates a modular AI-based classification component for predictive assessment of HR promotion cases. Multiple models were trained and evaluated on an imbalanced institutional dataset, including Logistic Regression, Random Forest, and their augmented variants using SMOTE, ENN, and Balanced Bagging techniques. The final selected model—Random Forest with Balanced Bagging and SMOTEENN—achieved an F2-Score of 0.97, Macro F1 of 0.67, and minority class precision of 0.31, demonstrating strong minority class sensitivity and overall balanced performance. Processing time was significantly reduced, with an end-to-end execution time of approximately 9.4 seconds for a standard two-page case batch and 3.3 seconds for a case batch of 1000 records using the developed AI model, in contrast to the average 3-day manual processing period. This proposed framework offers a novel approach to workflow efficiency improvements in potentially numerous sectors where similar document processing contexts are present.

Keywords: Data Extraction, Document Digitisation, Decision Support System, Artificial Intelligence, Unstructured Data.

ACKNOWLEDGEMENT

I would like to take this opportunity to express my gratitude and appreciation to my supervisor, Prof. Aaron Zimba for his guidance, patience and invaluable advice throughout this project. I also extend my heartfelt thanks to my lecturers—Dr. Bob Jere, and once again, Prof. Aaron Zimba—for their knowledge, insightful contributions, and dedicated supervision during my studies.

I would like to additionally express my thanks to the Teaching Service Commission of Zambia, for their support through the workings of this project.

THANK YOU.

DEDICATION

I would like to take this opportunity to express my heartfelt gratitude and appreciation to my parents, Mr. Wedex Ilunga and Mrs. Margaret M. Ilunga, for their unwavering love, support, and encouragement throughout the development of this work. My sincere thanks also go to my siblings, Mr. Chiwila Sharpe Ilunga, Ms. Chipampe Ilunga, and Ms. Musonda Ilunga, for their constant support.

I am in addition deeply grateful to Ms. Lomadinga M. Kasochi for her continual encouragement and invaluable support from the inception of this project and during my period of study at ZCAS University.

Lastly, I would like to thank my workmates and friends for their encouragement and support throughout the creation of this project. My sincere gratitude to them all.

LIST OF ABBREVIATIONS

Abbreviation	Full Term
ABSA	Aspect-Based Sentiment Analysis
AI	Artificial Intelligence
AUC	Area Under the Curve
BERT	Bidirectional Encoder Representations from Transformers
CNN	Convolutional Neural Network
CRM	Customer Relationship Management
CTC	Connectionist Temporal Classification
DSRM	Design Science Research Methodology
ELMo	Embeddings from Language Models
ENN	Edited Nearest Neighbour
FBSA	Fine-Grained Aspect-Based Sentiment Analysis
F1	Harmonic Mean of Precision and Recall
F2	Weighted Harmonic Mean prioritising Recall
GDPR	General Data Protection Regulation
GINA	Genetic Information Nondiscrimination Act
HOG	Histogram of Oriented Gradients
HR	Human Resources
HRIS	Human Resource Information System
HRM	Human Resource Management
IDMS	Intelligent Document Management System
IR	Information Retrieval
MD-RNN	Multidimensional Recurrent Neural Network
ML	Machine Learning
NLP	Natural Language Processing
NN	Neural Network
OCR	Optical Character Recognition
OM	Opinion Mining
PR AUC	Precision-Recall Area Under the Curve
RegEx	Regular Expressions
ROI	Region of Interest
RPA	Robotic Process Automation
SA	Sentiment Analysis
SMOTE	Synthetic Minority Over-sampling Technique
SMOTEENN	Combination of SMOTE and Edited Nearest Neighbour
SQL	Structured Query Language
SVM	Support Vector Machine
UI	User Interface
UML	Unified Modelling Language
[CLS]	Classification Token (used in BERT architecture)
[SEP]	Separator Token (used in BERT architecture)

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1 - INTRODUCTION

1.1 Background to the study

Human resource case processing is an essential element within any industry that employs a human workforce. It requires certified HR personnel to evaluate and resolve various employee-related cases. Timeliness and precision in case management are critical; delays or inaccuracies can have significant impacts on employees' careers and employer operations alike. HR case workers are thus tasked with understanding the complexities of their specialisation while handling sensitive information confidentially. This work includes recruitment, selection, and performance monitoring, processing of promotions, employee transfers, retirements and various other HR functions, each of which influences individual and organisational performance outcomes [32].

Technological advancements have reshaped HR functions, with Human Resource Information Systems (HRISs) improving the monitoring, documentation, and recording processes within organisations [33]. These systems have facilitated smoother integration of HR procedures into daily routines, underscoring their value within HR departments. Furthermore, technology now permeates nearly every HR activity, contributing to a range of HR management (HRM) functions and advancing operational efficiencies [34][35].

Artificial intelligence (AI) has emerged as a transformative force across global industries, noted for its ability to automate routine tasks, support data-driven decision-making, and enhance operational efficiency [36]. In HRM, AI technologies have expedited talent acquisition by simplifying candidate evaluation processes, thus reducing the burden on human personnel and improving organisational responsiveness to workforce demands [37]. AI-driven tools can also help identify and address employee performance issues proactively by offering training and coaching solutions, ultimately driving individual and organisational performance outcomes [38]. As AI systems continue to advance, the precision and speed at which they perform HR tasks may result in less reliance on human effort for certain functions, raising questions about potential job displacement and other socio-economic impacts [39].

Organisations are increasingly adopting AI driven solutions to automate and streamline HR functions, with future projections suggesting continual integration due to the efficiency, decision making and decision support benefits AI offers [40]. However, there are concerns over the ethical implications of AI, particularly regarding bias, judgement, job displacement and privacy concerns. Despite these challenges, AI demonstrates its potential to enhance HR

processes by refining decision accuracy, improving procedural efficiency, and supporting strategic HRM objectives [41][42]. AI systems have thus become integral in aligning HR functions with broader business goals, especially when HR management is included at the highest decision-making levels of an organisation [47].

Machine Learning (ML), a subset of AI- has garnered particular attention for its ability to enhanced automation through adaptive, data informed [49]. By facilitating real-time decision-making, ML models offer the potential to continuously refine HR functions in response to evolving organisational needs [2]. However, the effective use of AI in HRM depends on the technological awareness and skill levels of HR professionals; organisations must invest, therefore, in developing HR personnel understanding of AI to fully leverage its potential [43][44]. Additionally, the integration of AI into HRM workflow should be carefully managed to mitigate perceived risks, encouraging user acceptance and maximising operational benefits [45][46].

This thesis aims to address the challenges in manual, paper-based HR case processing systems commonly found within government HR departments. These traditional methods are prone to document deterioration, misplacement, and inaccuracies, which can delay decision-making and hinder case resolution. By developing an AI powered system combining automated image processing with custom, lexicon-based Aspect Based Sentiment Analysis (ABSA) for HR case document analysis, this project seeks to enable efficient, automated handling of HR cases with accuracy comparable to human judgement. In doing so, it contributes to the broader field of AI in HRM by demonstrating how AI-driven, adaptable systems can support scalable, reliable HR decision-making, further validating the alignment between AI advancements and organisational HR goals.

It is suggested in [48] that the integration of AI in HRM is particularly impactful when HR is involved in the strategic decision-making process at the highest levels of the organisation. AI systems serve as supportive tools, enhancing the effectiveness of HR functions by providing real-time insights to provide data driven recommendations, augmenting the skill and efficiency of HR professionals. This alignment between AI capabilities and organisational strategy enables HR professionals to contribute more meaningfully to broader business objectives. By streamlining HR tasks and aligning them with principal goals, to foster organisational growth.

1.2 Problem Statement

Human resource (HR) case processing is an integral part of organisational management. There are significant challenges in industries that rely on human labour. HR departments are responsible for evaluation and resolution of a wide range of employee related cases, where the timeliness and accuracy of case management are critical. Any delays or errors in processing can lead to detrimental impacts on employees' careers and the overall operational efficiency of an organisation. The management of HR cases requires specialist knowledge, strict confidentiality principles, and the ability to navigate complex nuanced situations. However, traditional manual, paper-based methods currently employed by many organisations often hinder efficiency. Such outdated systems are prone to document deterioration, misplacement, and inaccuracies, leading to delays in decision-making, duplication of work, general errors in case resolution, and reduction in operational efficiency.

Recent advancements in artificial intelligence (AI) have brought about new opportunities to address these challenges by automating and streamlining HR processes. However, AI's potential in HR case processing remains for the most part untapped, particularly within some government sectors where manual, paper-based systems are still prevalent. By leveraging AI through the use of OCR for automated HR case document processing, combined with text analysis, ABSA and AI. ABSA initially applied as a means to gauge sentiment expressed in printed text to provide decision recommendations on cases within printed text documents (batches). AI subsequently being applied by the development to of an optimised machine learning model to assess various case features to make case processing recommendations with stated degrees of certainty. to assess the context and sentiment of case-related content, this project seeks to address the inefficiencies and inaccuracies inherent in manual HR case processing systems. These technologies will enable more accurate and timely case resolution, reducing dependency on manual labour, and ensure that HR decisions are supported by reliable, data driven insights. By adopting AI-powered systems, organisations can optimise case processing, improve decision-making speed, and ensure a higher standard of service in HR departments. This research seeks to demonstrate how AI can facilitate scalable and reliable solutions for HR case processing, with the potential to refine the efficiency and accuracy of HR operations within governmental organisations.

1.3 Aim of the Study

The aim of this study is to develop an AI-driven system that enhances the efficiency and accuracy of HR case processing within organisations. Specifically, this system will leverage

OCR, custom ABSA and AI to automate the analysis and processing of HR case documents. The goal is to reduce the inefficiencies, errors, and delays inherent in traditional manual, paper-based systems, thereby improving the decision-making process within HR departments. This study will, furthermore, explore the utilization of machine learning techniques for the refinement of human resource case processing procedures with the integration of a more adaptive, data-driven model optimised for HR case processing.

1.4 Objectives

The objectives of this research are:

I. To identify and analyse the primary challenges in developing an AI-driven system for HR case processing.

II. To design and develop a system that integrates AI, OCR and custom lexicon-based SA to automate the processing of HR case documents.

III. To facilitate the accurate extraction of key information from scanned HR case documents using RegEx.

IV. To evaluate the performance of the developed system in terms of its accuracy, efficiency, and scalability, in comparison with traditional manual HR case processing methods.

V. To provide recommendations for the future refinement of integrated machine learning models, analysing methods with which the limitations of the technologies and their general augmentation may be achieved.

1.4.1 Research Questions:

The following are the research questions of this work:

1. What is the feasibility with which AI, OCR, RegEx and custom lexicon-based SA can be effectively integrated into a system to automate the analysis, processing and digitisation of HR case documents?
2. How do the metrics of accuracy, efficiency, and scalability of the proposed intelligent HR case document processing framework compare to current traditional manual processing methods?
3. How effective is the use of RegEx for accurately extracting key information from scanned HR case documents, and what are its identified limitations?

4. What are the primary challenges in developing an AI-driven system for HR case processing, and how can these challenges be mitigated?
5. How can the integrated machine learning models be optimised, and limitations mitigated or circumvented within the HR case processing framework and in addition, what are the key challenges and future research directions in this area?

1.5 Scope and Limitation

1.5.1 Introduction

This study focuses on automating the processing of Human Resource (HR) cases, specifically within environments that rely on traditional paper-based systems. The primary goal is to design and develop an AI-driven system incorporating Optical Character Recognition (OCR), custom lexicon-based Aspect Based Sentiment Analysis (ABSA) and Machine Learning Models to streamline HR case document analysis. This prototype system aims to address the inefficiencies, errors, and delays that can arise from manual processing, thus improving the overall accuracy and efficiency of HR decision-making.

The scope of the research includes the evaluation of the system's performance in terms of its speed, accuracy, and scalability within HR departments. This study will also examine the integration of machine learning models as an augmentative approach to the ABSA method, in order to make the system more adaptive to the evolving needs of organisations.

While the research will focus on a controlled, organisational context, it will not fully replace human involvement in decision-making processes for complex HR cases, but rather aim to augment and assist in the automation of routine tasks. This system, is perhaps best considered not a final solution, but a precursor to a greater functioning system that facilitates the interactive processing of digitised records across various organisational levels, thereby reducing upon delays, improving upon efficiency, and saving on time and workforce requirements on the processing of human resource cases and the digitisation of human resource case records.

Ethical considerations, such as data confidentiality and the potential for algorithmic bias will also be addressed, but the study will not focus extensively on these topics within its current scope. Due to confidentiality concerns and compliance with legislation governing the preservation of privacy, the data used shall undergo anonymisation. It is imperative to mention that no personal information on real entities shall be divulged or displayed in this research,

however, the research is intended for application in real world scenarios which do involve the processing of personal information.

1.5.2 Limitation 1 - Case Category Limitation

A key limitation of this study is that the developed prototype system will be tested using a narrow, focused set of HR case types and document categories. As a result, the system's performance may be optimal for the specific cases tested but may not be as effective in handling unseen case types or highly nuanced scenarios. While this restriction may impact the system's ability to generalise across all potential HR case scenarios, it ensures a directed approach to specific case types, allowing for detailed and thorough analysis and training within the selected areas.

1.5.3 Limitation 2 – Dataset Imbalance

A dataset of past human resource case records has been provisioned by the Teaching Service Commission of Zambia to allow the development of a Machine Learning model fit-for-purpose based on established precedent. However, the dataset possesses an imbalanced distribution of majority and minority class cases. This may be a limiting factor in the optimisation of the developed Machine Learning model, and as such: considerations must be taken to ensure the detrimental effects the skewing has on the developed module are reduced.

1.6 Significance of the Research

The proposed research aims to design and develop a framework that utilizes Optical Character Recognition (OCR) and Artificial Intelligence (AI) to address the inefficiencies and challenges currently faced by Human Resource (HR) departments within governmental contexts in case document processing and digitisation. The dependence on manual modes of processing and scrutinization of HR cases often leads to bottlenecks which lead to the accumulation of backlogs, and extended case processing times. By integrating OCR, AI and Text Processing technologies, this research seeks to streamline the scanning of documents, improve text character identification accuracy, digitise case data and automate decision-making processes, significantly reducing the need for human intervention [4][6].

The integration of AI and OCR in Human Resource Management (HRM) is essential in improving operational efficiency by automating tasks such as document handling, data extraction, digitisation and decision support. AI's ability to process vast amounts of data both swiftly and accurately makes it invaluable in HRM applications, particularly in document processing, recruitment, and performance management [1][20]. Through AI-driven solutions,

HR departments can not only speed up their case processing but also enhance the accuracy and reliability of decisions, leading to timely resolutions and more informed outcomes [24][25]. The inclusion of OCR further amplifies the capabilities of AI by enabling the automatic extraction of data from HR documents, reducing human error and ensuring higher precision in document processing.

Further, the proposed framework is expected to have seamless integration with existing HR management systems, offering a user-friendly interface with scalable architecture to adapt to different organizational contexts. The properties of scalability and adaptability ensure that the framework will remain effective as organisations evolve and grow, meeting the dynamic needs of HRM [8][9]. Security and compliance are also key concerns, particularly in the handling of sensitive HR data. The research emphasizes robust security measures to safeguard data privacy, ensuring that HR departments can trust the AI-powered system without compromising on confidentiality [23][28].

AI's role in automating decision-making processes is particularly relevant in recruitment, where AI systems can efficiently analyse resumes and identify the most suitable candidates based on predefined criteria, thus optimising HR functions by reducing time spent on manual screening and increasing the accuracy of hiring mechanisms [20][24]. Ethical implications, however, must be considered. Transparency in AI algorithms is essential to avoid biased decision-making, which could undermine fairness in recruitment and other HR processes [1][3][23]. Ensuring that AI systems remain free from training bias is crucial in maintaining trust and preventing reputational damage within organizations.

The use of fuzzy logic models in HRM systems can as well play a significant role in addressing the uncertainties and complexities often encountered in HR decision-making, such as degrees of aspects in performance evaluations and dispute resolutions. These models provide nuanced recommendations that improve HR decision quality, allowing systems to handle incomplete or ambiguous data [19][28]. AI's ability to handle such uncertainties while automating repetitive tasks leads to significant cost reductions and greater resource optimization within HR departments [24][25].

While AI and OCR technologies offer significant advancements, human involvement remains vital in ensuring ethical, effective operation of systems utilising these technologies. AI should augment human decision-making rather than replace it entirely, particularly in nuanced or complex HR cases where human judgment is crucial [30][31]. This research explores the

intricate balance between AI capabilities and human expertise, ensuring that systems driven by AI complement human judgment for more accurate and ethical HR decision-making and document processing.

Ultimately, this research contributes to the understanding of how AI and OCR can revolutionise HRM by improving efficiency, reducing costs, and enhancing the quality of decision-making. It provides a framework for HR departments with standardised document formats to leverage AI and OCR technologies, potentially offering significant economic and operational benefits while ensuring that sensitive data is managed with the utmost security and compliance. By enhancing case management procedures and HR processes, this research could offer a cost-efficient solution that mitigates the needs for expansion of departmental workforce expansion, especially in large institutions.

1.7 Preliminary Sections of the Project

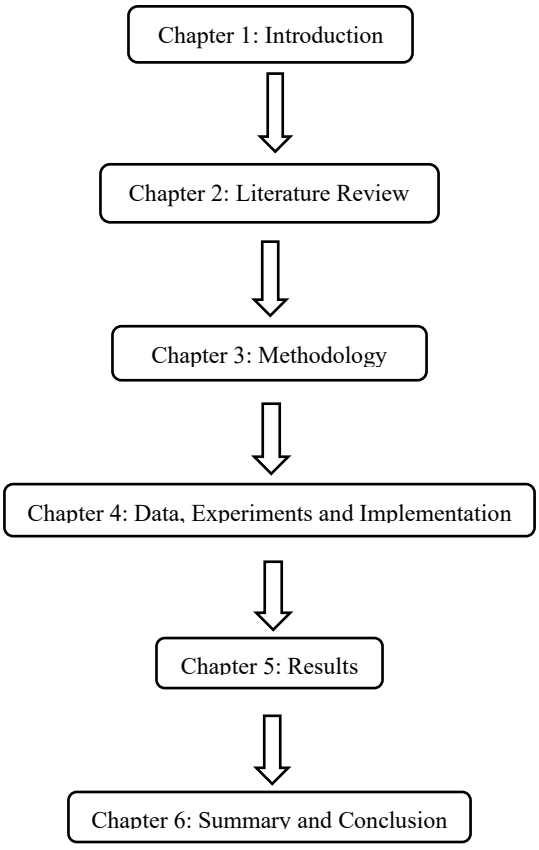


Figure 1. 1 Preliminary Sections of the Project

Conclusion

In summary, this report proposes the development and implementation of a framework for the automation of HR case document processing and digitisation using AI, OCR, SA and RegEx to pre-process images, extract text, identify sentiment and extract key data points, all to be digitised. This project is intended for deployment in governmental contexts, where HR documents follow standardised formats, structure and domain specific language.

CHAPTER 2 - LITERATURE REVIEW

2.1 Introduction

2.1.1 General Background

HRM involves various processes related to recruitment, performance management, employee development, transferences and promotions, and compliance. As organizations face increased challenges regarding the management of large data volumes in efficient manner, traditional HRM practices that rely heavily on manual processing become less viable. OCR and AI are two powerful technologies poised to transform HRM by automating and streamlining the processing of HR documents, improving efficiency, and reducing human error. OCR allows machines to read and digitise handwritten or scanned documents, while AI systems are capable of processing vast amounts of data to identify patterns, automate decisions, and optimize HR operations. The integration of these technologies promises to revolutionise HRM, particularly in areas where document processing is time-consuming, such as recruitment, performance evaluations, and compliance monitoring. One such area being in Governmental operating environments.

2.2 Broad literature review of the topic

2.2.1 Introduction

The broad literature review seeks to identify and elaborate on both relevant past developments in the field of AI and OCR with regards to HR document processing. It covers a variety of source material including published research articles, journal publications, conference papers and books from recent years and distant publications that are relevant or provide basis for current technologies. This mode of research allows for the establishment of strong knowledge base of the subject area and any theoretical and developments thereof.

2.2.2 On the HRM Domain: AI and OCR and Ethics

The utilisation of AI is more commonplace in the modern office. Enabling the solving of problems in dynamic and automated ways to improve on efficiency of operations, reducing upon costs and bolstering decision making and support processes, and allowing swift adaptation to the changing needs of a business. OCR when used in conjunction with HRM has been shown to possess the potential for similar benefits, including, strengthening decision making accuracy.

These technologies and others are driving forces in the field of task automation. If well applied, capable of transforming how case processing, and document management are executed. One important thought is that if technologies such as these are applied, the ethics and operational aspects of the development or integration must be well considered.

AI in HRM has been applied in the making of hiring decisions, freeing up time that would otherwise be dedicated to such processes by HR professionals. Large amounts of data, be it resumes or other documents can quickly be sifted through to identify what is needed and make a decision. This aspect of AI can greatly reduce workforce operational costs [4]. Further with regards, to OCR: OCR has the capability of expediting manual data entry by providing the option to digitise written, and printed text, leading to potential reductions in human error and subsequently leading to higher accuracy, by reducing the avenues through which it can occur [5].

Wherever employees are involved, there exist the processes and procedures of HRM. HRM personnel are vital in any industry that utilises a workforce. These personnel possess understanding of procedure and critical knowledge which is utilised when decisions must be made- nuanced, complex or otherwise. This all being done with the firm principles of confidentiality, as the work of HR frequently encroaches on sensitive personal information. This work includes tasks such as recruitment and selection of personnel, monitoring of personnel performance, the promotion, transference and retirement of personnel, and various other important functions that have bearing on organisational and employee success [32].

AI systems, specifically Machine Learning (ML) models are typically trained on vast amounts of data, this data may be data produced along the course of organisational operation. When the data with which a ML model is trained reflects bias or unfair patterns in case processing, the model shall exhibit the same inadequacies in its operation, greatly disadvantaging some personnel or prospective hires as the case of application may be [1]. It is important to ensure AI ML models are ethically built, subverting bias with intentional action, this in turn builds greater support and recognition of the new technology among employees and others as having merit to produce good [2].

The HR landscape has been reshaped by technological advancement, Human Resource Information Systems (HRISs) when applied have improved the monitoring, documentation and recording of organisational processes [33]. Systems such as these have been proven to smoothly facilitate the integration of routine HR procedures into the day-to-day activities of personnel, permeating nearly every HR and HRM activity and advancing operational efficiency [34][35].

Though AI and OCR promise many positives and advantages when correctly and ethically implemented, challenges still abound. A primary challenge is computational power availability, especially at scale and within large organisations where such scale is demanded.

Advancement requires the development of more sophisticated and demanding data pre-processing and optimisation technologies for use in the HR task domain. Constant and continuous advancement are pivotal for maintaining high performance and high achievement within organisations, and thus investment into the necessary computational architecture is of some pertinence [6][7]. Caution must be taken when implementing any system, including that built on AI technology, to ensure that users, i.e. HR personnel are trained, or the system may be underutilised and perform poorly.

There is increasing uptake of AI based systems by organisations in recent years, with future projections implying continued integration of the technology due to its benefits towards efficiency and decision making/support [40]. Ethical concerns on bias, potential job displacement, judgemental concerns, and privacy concerns still abide. Regardless of these concerns, AI demonstrates its usefulness and potential for positive transformation of HR processes to meet evolving HR objectives [41][42].

There exist Intelligent Document Management Systems (IDMS's) which leverage a combination of AI, OCR and Natural Language Processing (NLP) technologies. These IDMS systems have revolutionised document management in HRM contexts, providing leaps in processing operations such as automated text extraction and classification, allowing the digitisation of documents for posterity [8]. Blockchain technology when integrated into IDMS further provides improvements with regards to the maintenance of document history and integrity, which is of importance in HRM, especially when compliance with legislation is of importance, such as with regards to the European Union's GDPR (General Data Protection Regulations) [8].

ML has over time garnered a particular attention with regard to its ability to bolster operational automation through adaptive, and data driven models [49]. ML models incorporate real time decision making, offering continuous refinement towards organisational and HR departmental needs and objectives [2]. However, the effective use of AI in HRM depends on the technological awareness and skill levels of HR professionals; organizations must therefore invest in developing HR personnel's understanding of AI to fully leverage its potential [43][44]. Additionally, the integration of AI into HR workflows should be managed carefully to mitigate perceived risks, encouraging user acceptance and maximizing operational benefits [45][46].

AI within HRM and with regards to HR personnel and their tasks, must be applied in augmentative context rather than substitutive. HR professional, AI collaboration is imperative,

as human professionals may be able to handle more complex and nuanced matters, while AI systems may handle more routine and repetitive tasks, lending assistance to decision making processes (Merlin & Jayam, 2018).

HR tasks, such as candidate role suitability evaluation or subjective criteria-based employee performance assessment, benefit from the incorporation of fuzzy logic systems which provide means of dealing with degrees of membership of aspects that a case may revolve around [19][28]. This allows HR professionals to make more informed decisions, especially in situations where data is incomplete or unclear. In areas of work involving complex factual consideration as well as nuance, such as performance evaluations, and dispute resolutions, human oversight is an essential component [23].

Transparency and fairness are critical to ensure that AI systems do not perpetuate biases or reinforce inequalities. AI-driven decision-making can be prone to biases if not carefully monitored and regulated, particularly when systems are trained on historical data that may contain inherent prejudices [1]. With regards to ML model training, diverse training set sample sizes are essential to ensure wide applicability of a system as well as subversion of potential bias that may disadvantage demographics [2].

Complex Challenges and Opportunities in AI

AI has shown competence in complex judicial decision-making applications. While the sensitivity of the Judicial decision-making context is high, there seems to be the potential for AI applications to be integrated into this and other complex, sensitive and nuanced matters. Marking a significant step in the applications of AI thus far [30][31]. This perhaps alluding to AI taking up the more complex tasks in time with success. However, this may be a matter of great ethical concern.

Cost Reduction Benefits

AI and OCR may be beneficial in the way of resource and cost savings. Automation of repetitive tasks alleviates the burden of these tasks on HR professionals, allowing them to exercise their abilities in more complex applications where they may be more necessarily applied. This cost saving potential raises the attractiveness of AI solutions to organisations [24][25].

Technological Integration

AI and OCR are well integrated when the task at hand involves the extraction of information or digitisation of information from paper-based documents [4]. Deep learning may be applied to gain more nuanced insights that are found in complex cases such as dispute resolution cases in HRM [5]. This bolsters organisational process adaptability to changing business landscapes.

2.2.3 Challenges in HR-Specific AI Applications

HR activities tend to trend towards nuance. For instance, the assessment of employee quality is challenging in that it may be based on highly subjective performance metrics such as commonplace appraisal scores, within use even in governmental context. However, this metric has been frequently criticised on its validity, reliability and propensity for bias [51]. Should a ML model be trained on such information it may well end up inheriting the very same disadvantages.

2.2.3.1 Issues of Vendor Fragmentation and Software Compatibility

At present state many AI based HRM solutions exist, with vendors offering packages that are adept at specific tasks only, this leads to situations wherein an organisation may purchase numerous AI HRM systems, each to perform a specific task out of the scope of the other, potentially leading to incompatibility of data and a distinct lack of interoperability. Slowing the progressive development of AI HRM systems as a whole, due to the fragmentation of solutions. Many vendors undertake this practice, which in turn makes data integrations across varied functions challenging [52].

2.2.3.2 Data Limitations in HR Analytics

Data quality and data availability in volume are two major factors that data science in HR depends on. In situations in which data may be scarce, the data science discipline may struggle. Within private companies, certain case types, such as dismissals may be rare occurrences. Hence to develop predictive models around dismissals or dismissals with a very limited data set would result in an inadequate and inaccurate HR system [53][54]. With that in mind, government HR contexts tend to deal with vastly higher numbers of cases, which may lead to an overcoming of the hurdle of data sufficiency inadequacy, thus risk mitigation HR systems may be more feasible in the governmental context.

2.2.4 Algorithmic Management and Privacy Concerns

"Algorithmic management" is a means of employee monitoring and incentivisation applied typically to contractor employees to bolster employee productivity and conduct

[55][56]. The European Union's GDPR and "right to be forgotten" legislations provide frameworks for adequately managing employee data privacy concerns. It must be stipulated with clear boundary; what data is relevant to be considered when handling an employee case or considering a prospective employee for hire. The GDPR mandates that employers must responsibly handle employee data and delete said data upon request by the employee [57].

Some suggest, with the regard to addressing data privacy concerns, that the Genetic Information Nondiscrimination Act ought to be used as a model for employee privacy protection, from employer misuse [58]. Differential privacy techniques which allow algorithms to protect individual privacy are gaining popularity, as while they are able to protect individual privacy, they are still able to derive insights at population levels. Such techniques may be relevant towards the prevention of AI ML model discriminatory bias and practices in HR, leading to fairer work environments through positive HR practices [59].

Role of AI in Workforce Capabilities and Job Creation

AI is a powerful tool for the augmentation and extension of human abilities, through learning, sensory input and data processing, it is able to automate and assist with varied tasks [60]. Some critics fear replacement of the human workforce with AI, however there is a general consensus that despite fears of employee displacement, AI will transform workplace tasks and dynamics to more efficient processes [61][62][63].

The integration of AI based systems in HRM is more impactful when HR personnel are highly involved in strategic decision-making processes at higher levels. This way, the operation of AI solutions aligns well with organisational needs and allow for the streamlining of HR tasks to strategically drive performance outcomes and fostering organizational growth [48].

2.2.5 Sentiment Analysis Advancements: Methods and Approaches

Sentiment Analysis (SA) or Opinion Mining (OM), in ways is a combination of Information Retrieval (IR) [64], Natural Language Processing (NLP) [66], and ML [64]. Lexicon use and generation is a key aspect of SA, a Lexicon or Dictionary, being a collection of words, sentiments and score pairs.

Subjectivity and polarity detection are key activities within SA. Subjectivity detection is the classification of text as either objective or subjective [67]. Polarity detection works on text identified as subjective and determines whether it conveys a positive or negative sentiment overall. Various types and levels of text data can be subjected to this process [64].

State-of-the-art polarity detection methods often make use of static lexicons, such as the widely utilised SentiWordNet, or optionally modify lexicon scores to tailor to their unique application [68][69][70]. Static lexicons, tend to be reliable for general purposes. However, they lack the adaptability offered by dynamic lexicons, which posess elements of context and domain sensitivity and are capable of updating with new text terms or phrases [66][67]. However, dynamic lexicons may suffer lower accuracy rates due to lack of human supervision.

A study focused on fine-grained Aspect Based sentiment analysis (FBSA) revealed that the processes of stopword removal and lemmatization significantly raise result precision [66].

Traditional sentiment analysis is non granular, while sentiment can be identified, it is not presented with regards to which particular aspect it relates to. The approach may be inadequate when a text is focused on multiple aspects. ABSA as a technology bridges the gap enabling the identification of an entity's attributes and the sentiment expressed toward each [72].

ABSA made advancement during the SemEval 2014 (workshop), where Task 4 involved the use of ABSA on provided datasets. This task was not limited to evaluating sentiment but required assessing the sentiment for various aspects or attributes within the data, ensuring that each aspect contributed to the overall sentiment of the document [73].

Bidirectional encoder representations from transformers (BERT) is a pre-trained language model that considers word context from both (the) directions simultaneously [74]. This capability significantly improves performance in tasks such as sentiment analysis and question-answering systems [72]. Prior to BERT, models like ELMo also used bidirectional unsupervised learning to produce contextualized word representations [75]. However, BERT's approach, using bidirectional pre-training, allows it to extract richer context compared to models like ELMo that train on left and right contexts separately [75]. This is achieved through (the) masked language model (MLM) technique, which randomly masks words in a sentence during training, replacing them with a [MASK] token [72].

BERT's input text undergoes wordpiece tokenization, which breaks it into subword units [76]. In addition to these tokens, two special tokens are used: ([CLS]) for sentence classification at the beginning and ([SEP]) to separate sentences [72]. A language model learns context through techniques like word embeddings, which represent words as vectors in a vector space [77].

The SemEval 2016 Task 5 again focused on ABSA, expanding on ABSA's tasks. These tasks include: (1) Aspect (category) classification, identifying topics and aspects discussed in the text; (2) Opinion target expression (OTE), extracting linguistic expressions that refer to the reviewed entity for each entity-aspect pair; and (3) Sentiment polarity classification, determining the sentiment for each identified topic-aspect pair [78]. Despite ABSA's reliance on syntactic features, the natural variability of texts poses a challenge for traditional syntactic parsers. To address this, one proposed solution is sentiment sentence compression (Sent Comp), which pre-processes text before ABSA [79].

While ABSA can be more challenging with spontaneous (unstructured) texts, it is typically easier to implement in standardized documents where sentence structure remains consistent, despite minor variations in wording. A lexicon-based approach to sentiment analysis (SA), as discussed by [80], represents a technique requiring no supervision, utilising a dictionary of words that are assigned (positive or negative) sentiment values [84]. The approach can be adapted for domain specific contexts through customisation of the lexicon. For example, the word "bad" could have a positive sentiment in certain contexts if the lexicon reflects such nuances, allowing for flexibility in the sentiment analysis operation, especially when dealing with domain-specific language or slang.

2.2.6 Summary

In conclusion: AI and OCR's utilisation in HRM technologies offers significant benefit. Cost reduction, Efficiency improvements, and improved accuracy with regards to decision making and decision support are some notable benefits. Ethical considerations, however, must be taken into account when incorporating these technologies into workflows, to ensure that they operate in a fair, transparent manner, free from bias. ethical considerations such as fairness, transparency, and human oversight are prioritised. The future of AI in HRM lies in the symbiotic relationship between AI and human expertise, where AI augments HR professionals' capabilities rather than replacing them. By maintaining a careful balance between automation and human involvement, organizations can unlock the full potential of AI-driven HRM systems while safeguarding against the risks of bias and unfair outcomes.

Manual, paper-based document processing and storage is prone to many challenges. Documents deteriorate, becoming illegible, or may be misplaced, human involvement may also lead to typographical error and inaccuracies with regards to data entry. This project seeks to address these challenges and more in governmental contexts. By leveraging AI, Sentiment Analysis, OCR and Regular Expressions (RegEx), a system can be developed to automate

document processing in HR departments. The planned workflow begins with input of images, image processing with SA and RegEx, for case text analysis, key data point identification and structured data insertions into a designated data store. This project is intended to improve efficiency through the automated handling of HR cases with higher accuracy. For an AI based, scalable, decision support tool, that is tailored to and in alignment with organisation goals.

Facts must be retrieved by experienced human resource professionals through the consideration of personnel cases and their documents within. These documents, such as appointment letters, promotional letters, and criminal background checks, are traditionally processed manually to form a recommendation or submission that is passed through various stages of review until a final decision is made. Manual document processing can be time consuming. By implementing SA, the system seeks to streamline and augment decision-making and data digitisation by automating the analysis of case details, indicated within recommendation and submission documents which contain distilled information derived from the earlier considered personnel files. The system will in this context, thus serve as a "final authority," offering recommendations on final decision with a probability of decision certainty, thereby improving the efficiency of the decision-making process.

The next section shall involve critical review of the analysed research works, identifying their points of alignment and consensus as well as highlighting differing opinions in order to provide an impartial view of the considered research material.

2.3 Critical Review of Related Works

2.3.1 Introduction

Building upon (the) information provided in the broad literature review this section provides a review of the literature from a neutral critical point of view in order to glean insights on both alignments and challenges posed by the differences of opinion in the research.

2.3.2 AI and HRM Integration

The integration of Artificial Intelligence (AI) into Human Resource Management (HRM) has been extensively discussed in academic literature, with (a) growing focus on improving efficiency, decision-making, and operational outcomes. This section critically reviews key studies to contextualise the unique contributions of this project to the field.

References [1] and [2] explore the transformative role of AI in HRM. Reference [1] highlights how AI adoption enhances HR processes, increasing operational efficiency and decision accuracy, while reference [2] investigates ethical decision-making facilitated by

algorithmic approaches. Despite their comprehensive analyses, these works do not delve into AI-driven HR case document analysis, a gap addressed by this project through its integration of Optical Character Recognition (OCR) and Aspect Based Sentiment Analysis (ABSA).

2.3.3 Aspect Based Sentiment Analysis (ABSA)

The potential of ABSA in HRM is underexplored, even though its applications for sentiment and textual analysis are promising. Reference [39] examines AI's role in various HR functions but do not explore sentiment analysis in HR-specific contexts. Similarly, reference [36] focuses on decision-making in organisations using AI but neglect the granular textual analysis required in HR case processing. By integrating ABSA with OCR, this project uniquely addresses these gaps, offering a framework for efficient and precise case handling.

2.3.4 OCR Technology

Reference [9] provides a detailed review of OCR technologies, highlighting their efficacy in automating text extraction from physical documents. Their findings underscore OCR's utility in enhancing document digitisation, a foundation for this project's objectives. However, their study does not explore the combination of OCR with AI-driven tools. This integration is a critical innovation of the current project, enabling a more comprehensive analysis of HR case documents.

2.3.5 Ethical Implications of AI

Ethical considerations in AI deployment remain a significant concern [2] [28]. These studies emphasise the importance of transparency, fairness, and bias mitigation in AI systems, particularly in sensitive domains such as HRM. The ethical challenges discussed in these works reinforce the importance of integrating robust safeguards within the proposed AI-powered system, ensuring that it aligns with ethical best practices in data security and algorithmic accountability.

2.3.6 The Strategic use of AI in HR

Reference [48] and [40] explore the strategic advantages of AI integration in HRM, including its ability to improve decision-making and operational efficiencies. Reference [48] examines AI's role in streamlining HR processes, while reference [40] focuses on user acceptance of AI-integrated Customer Relationship Management (CRM) systems. Although these studies provide valuable insights, they do not address the specific challenges of HR case processing in government sectors, which this project uniquely aims to resolve.

2.3.7 Identified Gaps

While existing literature underscores the benefits of AI, significant gaps remain in applying AI, OCR, ABSA and Artificial Intelligence to HR case processing. Few studies have explored how these technologies can be combined to enhance HRM functions effectively. This project addresses these gaps by developing a system that integrates OCR for text extraction with ABSA and AI for sentiment-based decision support. Additionally, it contributes to the broader field by addressing ethical challenges, such as data privacy and algorithmic bias, ensuring a secure and trustworthy AI application.

The next section involves a comparison of this project with other similar project and projects that feature similar functionality or context in part or whole to derive insights into potential solutions or challenges that may be encountered.

2.4 Comparison with related works

2.4.1 Introduction:

The previous section involved a critical review of the considered literature in an effort to identify various aspects of coalescence and dissonance across the considered papers, in conclusion of the chapter was a discussion on the identified gaps within which the solution of this project may operate. This section provides a comparison of existing works in relation to this project's aim of automating HR case processing using OCR, ABSA, and AI while addressing the inefficiencies of traditional HR case processing systems.

The integration of Artificial Intelligence (AI) into Human Resource Management (HRM) has been widely discussed in academic literature, showcasing its potential to revolutionise HR processes.

2.4.2 Comparison Table presenting OCR Model Accuracy across various implemented and experimental technologies

Table 2. 1 OCR Model Accuracy Comparison

Source No.	Reference	OCR Model	OCR Accuracy	Text Type	Dataset/Document Type	Notes
[8]	M. Pandey, M. Arora, S. Arora, C. Goyal, V. K. Gera, and H. Yadav, "AI-based integrated approach for the development of intelligent document management system (IDMS)," *Proc. Comput. Sci.*, vol. 229, pp. 725–736, 2023, doi: 10.1016/j.procs.2023.12.127.	EasyOCR	87.53%	Printed	Medical invoices from a curated subset of 70 images collected with hospital and patient consent	Used in IDMS framework
[8]	Same as above	PyTesseract + OpenCV + NLP	97%	Printed	Medical invoices (same dataset)	Higher performance via preprocessing
[14]	S. D. Connell and A. K. Jain, "Template-based online character recognition," *Pattern Recognit.*, vol. 34, no. 1, pp. 1–14, 2001.	Template-based + Decision Trees	86%	Handwritten (Online)	Character samples showing stylistic handwriting variation	Recognises handwriting styles using decision trees
[85]	D. J. Burr, "Designing a handwriting reader," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-5, no. 5, pp. 554–559, 1983, doi: 10.1109/TPAMI.1983.4767435.	Constrained single-stroke input + dictionary aid	90% → ~100% (with dictionary)	Handwritten (Segmented)	Characters written using single-stroke enforcement	Dictionary-based correction significantly improved recognition
[86]	B. P. Berman and R. J. Fateman, "Optical character recognition for typeset mathematics," in *Proc. ACM Conf. Document Processing Systems*, Jan. 1994, doi: 10.1145/190347.190438.	Commercial OCR systems (unspecified)	~99% → ≤10% on math content	Printed (Math)	Typeset mathematical expressions	Recognition drops due to layout variation and symbol complexity
[87]	U. Garain, "Identification of mathematical expressions in document images," in *Proc. Int. Conf. Document Analysis and Recognition*, Barcelona, 2009, pp. 1340–1344.	Feature-based mathematical OCR	97% (text with math), 95% (embedded), 97% (displayed)	Printed (Math)	Document images with embedded/displayed expressions	Symbol adjacency statistics boost recognition

Source No.	Reference	OCR Model	OCR Accuracy	Text Type	Dataset/Document Type	Notes
[88]	A. Graves and J. Schmidhuber, "Offline handwriting recognition with multidimensional recurrent neural networks," in *Proc. 22nd Int. Conf. Neural Information Processing Systems (NIPS)*, 2008, pp. 545–552.	**MD-RNN** + CTC	91.4% (Arabic handwriting)	Handwritten (Offline)	Arabic handwriting competition dataset	No language-specific preprocessing; outperformed all other systems
[89]	C. Malon, S. Uchida, and M. Suzuki, "Mathematical symbol recognition with support vector machines," *Pattern Recognit. Lett.*, vol. 29, no. 9, pp. 1326–1332, 2008, doi: 10.1016/j.patrec.2008.02.005.	Infty OCR + SVM filtering	96.10% → 97.70%	Printed (Math Symbols)	Test set of mathematical symbols from Infty system	SVM filtering of confusion clusters cut errors by 41%
[90]	N. A. Jebril, H. R. Al-Zoubi, and Q. A. Al-Haija, "Recognition of handwritten Arabic characters using histograms of oriented gradient (HOG)," *Pattern Recognit. Image Anal.*, vol. 28, no. 2, pp. 321–345, 2018.	HOG + SVM	99%	Handwritten (Offline)	Names of Jordanian cities, towns, and villages	Feature-based recognition of Arabic characters
[90]	Same as above	Multichannel NN (segmentation) + **CNN** (recognition)	94.38%	Printed	18pt machine-printed font	End-to-end system for machine print OCR
[91]	Lenz Furrer and Martin Volk, "Reducing OCR errors in gothic-script documents," in *Proc. Workshop on Lang. Tech. for Digital Humanities and Cultural Heritage*, 2011, pp. 97–103.	OCR with lexicons and n-grams	96.72% → 98.36%	Handwritten (Gothic)	Gothic script texts	Uses complementary resources (e.g., German dictionary, local place names) for error correction

2.4.3 AI for increased efficiency in HRM

Reference [1] highlights the transformative role of AI in automating repetitive HR tasks and enhancing decision-making processes. Similarly, reference [24] discusses the evolution of electronic HRM (e-HRM) systems over the past four decades, emphasising their role in

improving administrative efficiency. However, both works primarily address automation in areas such as recruitment and employee performance evaluation, lacking specific focus on HR case processing. By contrast, this project extends these insights into the domain of government HR settings, targeting document-specific challenges through OCR, ABSA and AI technologies.

2.4.4 Document Digitisation and OCR integration

Reference [9] provides a comprehensive review of OCR applications in digitising handwritten and printed documents, highlighting its utility in administrative processes. This aligns closely with the foundational role of OCR in this project, where document digitisation is a critical precursor to textual and sentiment analysis. Reference [4] explores OCR's integration with Robotic Process Automation (RPA), offering a model relevant to this project's goal of automating HR document workflows. These studies provide a strong technological foundation, but they do not delve into the sentiment analysis component addressed by this project.

2.4.5 Sentiment Analysis for Decision Support Systems

Reference [40] explores AI-driven tools that enhance organisational decision-making through data analysis. Although their focus is on customer relationship management, their methodologies align with this project's implementation of ABSA to assess the context and sentiment of HR documents. Reference [39] also examines sentiment analysis as a tool for improving decision-making accuracy, further underscoring the relevance of this approach to the project's objectives. These studies validate the project's emphasis on contextual text analysis as a means of improving the timeliness and accuracy of HR decisions.

2.4.6 Ethical Considerations in AI solution adoption

Reference [2] and [28] address ethical concerns in AI systems, including algorithmic bias, transparency, and data privacy. This project incorporates these considerations by ensuring that AI tools are deployed ethically, particularly when handling sensitive HR data. Bhave et al. (2020) emphasise the importance of privacy in workplace AI applications, reinforcing the project's focus on secure data handling and compliance with ethical standards.

2.4.7 Human-AI Collaboration in HRM

Reference [18] advocates for a collaborative relationship between human decision-makers and AI systems, where AI complements rather than replaces human judgment. Similarly, reference [45] discusses the interplay between AI technologies and human roles in the workplace, suggesting that AI can enhance but not substitute complex decision-making processes. This project aligns with these perspectives by designing an AI system that supports

HR professionals in processing cases, ensuring that human oversight remains integral for nuanced or complex scenarios. By allowing AI to handle routine tasks and support data-driven insights, HR professionals can focus their efforts on complex cases requiring contextual understanding and strategic decision-making.

2.4.8 Advanced Applications of AI in HR

Reference [36] demonstrates AI's capability to enable real-time decision-making in healthcare HR functions, a parallel to this project's goal of improving decision timeliness in government HR settings. Reference [19] explores the broader potential of intelligent systems in automating complex tasks, providing a theoretical basis for this project's application of adaptive AI models in HR workflows. These insights highlight the scalability and relevance of this project in advancing operational efficiency within HRM.

2.4.9 The Unique Contributions of this Project in Alignment with Identified Gaps

While the reviewed studies offer valuable insights, they leave significant gaps, particularly in addressing HR case processing in government settings. This project stands out by combining OCR and other technologies for document digitisation with ABSA for sentiment and contextual analysis as well as AI for decision support, a novel integration aimed at automating HR case handling with accuracy comparable to human judgment. Furthermore, the project addresses specific challenges such as document deterioration and inefficiencies inherent in manual processing, setting a benchmark for future research in HR automation.

2.5 Conceptual Framework

2.5.1 Introduction:

The previous section involved a critical consideration of the research material within the subject area, in order to identify challenges, points of alignment within the research and the project and gaps within implementation of the researched technologies in HRM. This section shall highlight the conceptual framework that shall be used for the development of the proposed solution.

The proposed project integrates theories and practices from artificial intelligence (AI), optical character recognition (OCR), Aspect Based sentiment analysis (ABSA) and Text Processing to address inefficiencies in HR case management. At its core, the framework conceptualises the coalescence between technological advancements and human resource management (HRM), seeking to optimise decision-making processes while maintaining ethical integrity.

2.5.2 Key Components of the Framework

2.5.2.1 Artificial Intelligence in HRM

AI serves as the backbone of the project, enabling the automation of data processing and decision support. Building on studies like references [17] and [23], the framework leverages AI to improve the accuracy and efficiency of HR functions, such as document analysis and case management. The theoretical foundation rests on the principle of machine learning (ML) as an adaptive tool capable of recognising patterns and making predictions to aid human judgment [18].

2.5.2.2 Optical Character Recognition (OCR) for Document Processing

The OCR module plays a crucial role in digitising physical HR documents, converting them into machine-readable text. Grounded in the technological advancements described by reference [10] and [9], this component ensures that manual, paper-based workflows are transformed into efficient, automated processes. The integration of OCR directly aligns with the project's goal to reduce document deterioration, misplacement, and inaccuracies.

2.5.2.3 Aspect Based Sentiment Analysis (ABSA)

ABSA is employed to evaluate the sentiments and contexts within HR case documents, a methodology inspired by computational linguistics. Reference [19] and [40] provide theoretical support for the use of intelligent systems to handle ambiguous and nuanced textual data. ABSA's role in identifying sentiment polarity and extracting specific aspects from textual data supports the decision-making process by offering precise insights into employee-related issues.

2.5.2.4 Text Processing (Lemmatisation, Stop-Word Removal and RegEx Pattern Matching)

Python's NLTK, and OpenCV modules shall be utilised to perform initial text extraction. The text shall then be further processed using text pre-processing technologies for lemmatization and stop-word removal for preparation of the data for ABSA or AI model-based analysis. Utilizing RegEx matching on un-processed or partially processed text then allows for key data elements to be extracted from the text by leveraging the text pattern consistency of governmental HR documents. Thus, allowing for digitisation of the records in addition to the processing of the text using ABSA to enable automation towards decision support.

2.5.2.5 Human-AI Collaboration

A hybrid approach underpins the conceptual framework, emphasising collaboration between human expertise and AI capabilities. Studies like [30] and [28] underscore the

importance of balancing automation with human judgment, particularly in complex, high-stakes decisions. By involving HR professionals in the final decision-making stages, the framework ensures ethical and contextual appropriateness.

2.5.2.6 Ethical Considerations and Transparency

The framework integrates ethical guidelines to mitigate risks of algorithmic bias and ensure privacy and fairness, aligning with concerns raised by references [42] and [28]. Transparency in AI algorithms and decision-making processes is paramount, aiming to foster trust and adoption among HR practitioners

2.6 Conceptual Model

2.6.1 Framework Phases

2.6.1.1 Data Input

HR case documents are digitised using OCR technology, ensuring accuracy in data capture and storage.

2.6.1.2 Data Analysis

Sentiment analysis and RegEx text mining techniques, SA, and AI extract relevant insights from digitised documents, focusing on specific employee-related aspects.

2.6.1.3 Data Extraction and Storage

RegEx pattern matching shall be utilised to perform extraction of text keywords, for storage, allowing digitisation of HR case material.

2.6.1.4 Decision Support

Insights derived from AI tools are presented in a user-friendly interface for HR professionals, enabling informed decision-making.

2.6.1.5 Feedback and Adaptation

Continuous user feedback refines the AI system, enhancing its accuracy and adaptability for future cases.

2.6.2 Proposed model

The proposed model for automating HR case processing combines key technologies—Optical Character Recognition (OCR), Aspect Based Sentiment Analysis (ABSA), and Artificial Intelligence (AI)—to streamline, analyse, and support decision-making in HR processes. This model focuses on improving efficiency, accuracy, and scalability in HR case management while integrating ethical guidelines to ensure fairness and privacy.

2.6.3 Components of the Proposed Model

2.6.3.1 Document Digitisation (OCR Integration)

The model begins with an OCR module for transforming paper-based HR documents into machine-readable formats. Leveraging proven OCR frameworks, such as those described by reference [10] and [9], the system captures text and metadata, ensuring high precision in character recognition. This component addresses issues of document misplacement, deterioration, and accessibility.

2.6.3.2 Text Analysis and Sentiment Assessment (ABSA)

Using ABSA, the system processes digitised text to extract aspects and associated sentiments within HR case documents. This stage is critical for evaluating nuanced content, such as employee grievances or performance reports. Grounded in adaptive algorithms [19], this module enhances the understanding of employee sentiment and provides actionable insights.

2.6.3.3 Data Extraction and Storage

Using RegEx, the system shall extract from the unprocessed text or partially processed text, key factors that make up elements of a human resource case, then passing the pattern matched values into a structured data structure to complete the record digitisation process.

2.6.3.4 AI-Driven Decision Support

Machine learning algorithms are applied to deliver predictive analytics and assist HR professionals in making data-driven decisions. AI systems, as described by references [18] and [20], allow for continuous learning and adaptation, improving the system's recommendations over time.

2.6.3.5 Human-AI Collaboration Interface

A user-friendly interface is designed for HR professionals to review AI-derived insights and make final decisions. Drawing from the emphasis on human-AI collaboration in reference [30], the model ensures that automation supports but does not replace human judgment, especially in complex cases.

2.6.3.6 Feedback and Model Refinement

A feedback loop is incorporated to refine the model continuously. User inputs and case outcomes are used to retrain machine learning algorithms, enhancing the system's effectiveness for future cases.

2.6.4 Workflow of the Model

2.6.4.1 **Input Stage**: HR case documents are scanned and converted into structured text using OCR.

2.6.4.2 **Pre-processing Stage**: The digitised text undergoes cleaning, tokenisation, and sentiment tagging.

2.6.4.3 **Analysis Stage**: ABSA identifies key aspects, such as employee performance metrics, and evaluates sentiment polarity.

2.6.4.4 **Data Extraction and Storage Stage**: Utilizing a set of prepared RegEx matching patterns, text is processed to extract key aspects of a case for storage into a structured data store.

2.6.4.5 **Decision Support Stage**: Insights are visualised for HR managers to review, and predictions are provided for potential case resolutions. This is accomplished through either ABSA or AI model-based analysis.

2.6.4.6 **Output Stage**: Finalised decisions are recorded on digitised data structure and integrated into organisational HR systems, with an option for feedback.

2.6.5 Advantages of the Proposed Model

2.6.5.1 **Efficiency**: Automates repetitive tasks, such as document processing and initial sentiment assessment.

2.6.5.2 **Scalability**: Capable of handling a growing volume of HR cases with minimal increase in resource allocation.

2.6.5.3 **Digitisation**: Provides opportunities for digitizing documents in large quantities provided a robust collection of patterns are created for data extraction and documents scanned follow a standard form.

2.6.5.4 **Accuracy**: Reduces human error in data extraction and case analysis through AI-enhanced analysis.

2.6.5.5 Ethical Governance: Incorporates mechanisms to prevent algorithmic bias and uphold data privacy by focusing on the relevant aspects of the case exclusively.

2.6.6 Future Extensions

The model can be further enhanced by:

- Integrating advanced machine learning techniques for adaptive decision-making.

- Expanding the dataset to include diverse HR cases for better generalisability.

- Employing fuzzy logic models for nuanced recommendations in ambiguous scenarios [28].

This proposed model addresses the challenges of manual HR case management by introducing a structured, technology-driven approach that maintains human oversight while leveraging the power of AI.

2.7 Chapter Summary

This chapter is a review of varied published works of informative, reflective and innovational relevance to the subject matter. It contains cross examinations of literature wherein alignments and deviations are identified and highlighted. Publications were also assessed in order to identify research gaps and opportunities in which novel technologies or novel applications of technologies may be leveraged.

Additionally highlighted within this chapter are details of the proposed framework and model for execution of this project research. The next chapter shall focus on applied methodologies within the project.

CHAPTER 3 - METHODOLOGY

3.1 Research Design

3.1.1 Introduction

This research aims to derive insights into the current state and future prospects of utilising Artificial Intelligence in document processing, with a particular focus on HR case processing. The study explores technologies such as Image Processing, Text Analysis, and Artificial Intelligence, alongside their evolution, to identify viable, replicable solutions aimed at improving operational efficiency and reducing reliance on paper-based document storage systems when unnecessary.

3.1.2 Design Science Research Methodology

This research adopts the Design Science Research Methodology (DSRM) [93] for the iterative development of the proposed framework.

3.1.2.1 Problem Identification

The solution is proposed to resolve the challenges encountered within manual paper based, Human Resource Case processing methods employed by many organisations. The solution utilises AI, OCR and Text Analytics to automate [9] and improve the efficiency of the HR case processing[19][40]. The complete problem statement is detailed within Chapter One of the research.

3.1.2.2 Objectives of the Solution

This research focuses on developing an AI-based system for automating HR case processing, using OCR, RegEx, lexicon-based sentiment analysis, and a trained AI classification model. It also evaluates the system's performance and explores future use of machine learning. The detailed list of objectives can be found in Chapter One of the research.

3.1.2.3 Design and Development

The design and development details of the solution are detailed in Chapters Three and Four of this research, encompassing both the theoretical-conceptual design, as well as the development of the functional prototype, respectively.

3.1.2.4 Demonstration

A demonstration of various modules of the solution is outlined within this research and shall be detailed with the use of solution images and code snippets as well as input/output results. The detailed information is included within Chapter Four of this work.

3.1.2.5 Evaluation

The evaluation of the developed solution was conducted, revealing insights into its effectiveness towards addressing the challenges earlier stipulated within the HR Case Processing domain. The evaluation is inclusive of performance statistics and comparisons with alternate methods. The detailed Evaluation is elaborated on within Chapter Five of this research.

3.1.2.6 Results

The presentation of produced results is available in Chapter Five of the research and includes detailed information and interpretations of achieved metrics of the developed solution as well as a review of the achievements of the solution developed against the set objectives.

3.1.3 Research Coverage

This research analyses a wide array of academic papers addressing various aspects relevant to the project. Selection criteria prioritised **recency**, focusing on papers published within the last five years, alongside **relevance**, targeting works that provide foundational insights or directly applicable information to the research focus. While the primary emphasis was on recent publications to capture the current state of the field, allowances were made for older yet pivotal works that introduce constant or enduring concepts crucial to understanding the evolution of the technologies under discussion. This balanced approach ensured a comprehensive understanding, merging up-to-date findings with essential historical perspectives.

The cited works include journal articles, technical papers, systematic literature reviews, books, conference proceedings, and workshop papers. This diverse literature set provided valuable perspectives and in-depth analysis, contributing significantly to the conceptualisation and development of the proposed solution.

Non-statistical data has been extracted from the selected research materials for inclusion within this report. Due to the project possessing a practical aspect in addition to the theoretical

work, additional data shall be obtained from the performance and functionality evaluations of the developed solution. Various aspects shall be measured for performance with consideration as well placed on available compute power of the unit within which the solution is run.

In addition to technical system data and data extracted from considered research publications, and simulated data shall also be used for the testing and development of the solution, along with a small set of anonymised real world case data. This data shall be used to test and ensure replicability of test results with the goal to ensure consistency and scalability.

3.2 Research Data and Datasets

3.2.1 Introduction

This section outlines the data sources and the systematic process used to select the research articles and datasets integral to this study. The research focused on the domains of AI, Image Processing, OCR, Text Processing and Sentiment Analysis with a specific focus on Aspect-Based Sentiment Analysis, particularly within the context of document processing in governmental Human Resource (HR) workflows.

3.2.2 Data Collection Methods

The sources of data include journal articles, conference proceedings, technical papers, and systematic literature reviews, selected for their **recency** and **relevance**. Recent works, such as those exploring cutting-edge AI techniques [72][74], were prioritised to align with the state-of-the-art advancements in artificial intelligence and document processing. Foundational studies, including references [18] and [19], were also incorporated to provide a robust theoretical basis.

The strategy included criteria for recency, generally selecting works published within the last five years, while also allowing older yet critical papers with enduring relevance to the field. For instance, foundational works in OCR [9][10] and sentiment analysis [78][79] were analysed for their contributions to the methodologies applied within this project. Relevance was determined based on each source's focus on technologies such as BERT [74], transformer models, and lexicon-based methods [80].

The diversity of sources ensures a holistic perspective on AI's applications in human resource management, including the ethical considerations raised by reference [2] and the

automation challenges identified by reference [1]. These insights collectively contribute to the development of a replicable, efficient, and ethically aware system for HR case processing.

3.2.3 Data Sources

The data utilised in the development of this work includes Primary Source Data. This data is further segmented into collected data from two major sources, including blinded data provided for the sake for this research by the Teaching Service Commission of Zambia, and other attributed online sourced datasets, from major data repositories including Kaggle and GitHub. Synthetic (Secondary) Data was not utilised in the development of this work. The rationale being that the work intended to produce a functional model for practical application that would correlate directly to the current state of the domain and the practices within it (i.e. established practices, procedures and patterns within the Humar Resource Case Document Domain within Governmental Institutions.

3.2.4 Data Source Functionality

The data sources utilised may be split into two categories, each with distinct purpose. The categories of collected data are:

I. *Primary Institution-Provided Data*, provided by the Teaching Service Commission of Zambia for the purpose of solution development and practical correlation to the work procedures and patterns of the researched domain; and
II. *Primary Sourced Repository Data*, obtained from Kaggle [94][95] and GitHub [96], comprising employee performance and promotion datasets used to establish a baseline (control) model for comparison with the solution developed using data from Source I.

Table 3. 1 Summary of Data Sources and Functional Roles

Data Source Category	Source Description	Source Reference	Function / Role in Study
Primary Institution-Provided Data	Blinded HR case data provided by the Teaching Service Commission of Zambia.	Not publicly accessible; internally sourced.	Used to develop the core solution. Data reflects real patterns, procedures, and decision processes within a government HR document handling context.
Primary Sourced Repository Data	Online open-source datasets related to employee promotion and performance. Includes: • HR Classification for Promotion; • HR Analytics; and • Employee Promotion Prediction.	[94], [95], [96]	Used to develop a baseline (control) model. Enables comparative evaluation against the solution built with institution-provided data.
Synthetic / Secondary Data	Not used in this study.	N/A	Deliberately excluded to ensure relevance to real-world application and alignment with institutional practices and domain-specific decision-making processes.

3.2.5 Data Source Technical Details

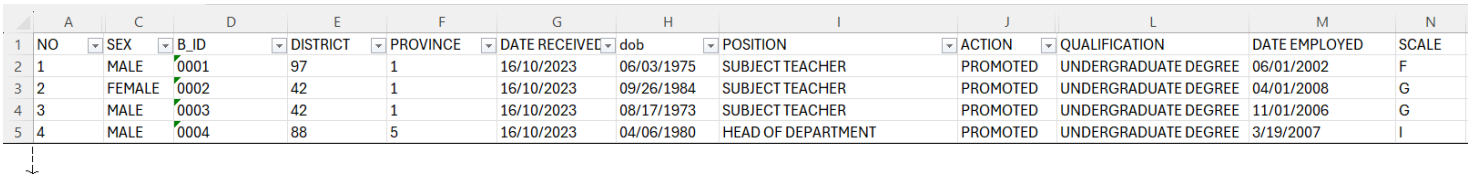
Source I Data:

The Data utilised consists of two types:

a) Scanned document images detailing case information for development of OCR, NLP, ABSA modules; and
b) Case data table detailing processed case records for AI Model Development.
    - The case data table consists of **11 Data Columns**, inclusive of a singular classification column, and features a **sample size of 1312 Rows**. The "Action" column shall be used as the target classification column and includes a range of two options: namely "Promoted" and "Not Promoted", replaceable for numerical binary style classification with "1" and "0" respectively.

- **11** Data columns are to be utilised for the development of an AI classification model. It is intended that derived metrics (such as Age) may as well be used though not explicitly found within the available columns. Four model types shall be considered and assessed to identify the model of the highest performance. Algorithms for the models shall be developed using a local operating environment with the Python programming language. These models shall include:
    - Logistic Regression Model;
    - Logistic Regression Model with SMOTE;
    - RandomForest Model;
    - RandomForest Model with SMOTE;
    - Random Forest Model with SMOTE-ENN;
    - Balanced Bagging Random Forest Model with SMOTE-ENN; and
    - XG Boost Model.

- A sample image of the data set is shown below:



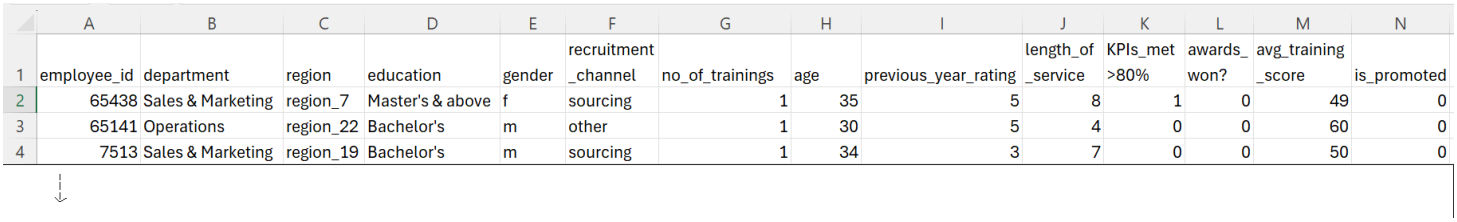
	A	C	D	E	F	G	H	I	J	L	M	N
1	NO	SEX	B_ID	DISTRICT	PROVINCE	DATE RECEIVED	dob	POSITION	ACTION	QUALIFICATION	DATE EMPLOYED	SCALE
2	1	MALE	0001	97	1	16/10/2023	06/03/1975	SUBJECT TEACHER	PROMOTED	UNDERGRADUATE DEGREE	06/01/2002	F
3	2	FEMALE	0002	42	1	16/10/2023	09/26/1984	SUBJECT TEACHER	PROMOTED	UNDERGRADUATE DEGREE	04/01/2008	G
4	3	MALE	0003	42	1	16/10/2023	08/17/1973	SUBJECT TEACHER	PROMOTED	UNDERGRADUATE DEGREE	11/01/2006	G
5	4	MALE	0004	88	5	16/10/2023	04/06/1980	HEAD OF DEPARTMENT	PROMOTED	UNDERGRADUATE DEGREE	3/19/2007	I

Figure 3. 1 Source I Data Set Sample

Source II Data:

- This data consisted of three open-source datasets [94][95][96] from online repositories. Each to be used for the development of models for the purpose of establishing a benchmark of ideal performance. The three data sets shall each be used, with the intention of the best performing model according to produced metrics being used as a baseline to gauge solution quality. The column is_promoted, shall be used as target column for the development of the classification model. A Logistic Regression based model is proposed for development, utilizing all 13 data points (exclusive of unique ID columns).

- A sample image of one of these datasets [96] is shown below, which features **13 Data Columns**, and a **sample size of 54,808 Rows**:



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	employee_id	department	region	education	gender	recruitment _channel	no_of_trainings	age	previous_year_rating	length_of _service	KPIs_met >80%	awards_ won?	avg_training _score	is_promoted
2	65438	Sales & Marketing	region_7	Master's & above	f	sourcing	1	35	5	8	1	0	49	0
3	65141	Operations	region_22	Bachelor's	m	other	1	30	5	4	0	0	60	0
4	7513	Sales & Marketing	region_19	Bachelor's	m	sourcing	1	34	3	7	0	0	50	0

Figure 3. 2 Source II Data Set Sample

3.2.6 Data Pre-Processing

3.2.6.1 Source I (a) Pre-processing

The Data is pre-processed for the optimisation of text extraction through OCR and NLP techniques. Some steps are pre-cursors to other pre-processing steps and must be completed prior in order to achieve the desired pre-processed state. The pre-processing of the images shall consist of the following steps:

- Grayscale, Black and White Conversion, Binarization: Precursors to further pre-processing steps, as well as for the simplification of image data for NLP processing.
- Noise Removal: Removes visually interfering elements that may potentially hinder OCR accuracy.
- Font Adjustments: Optionally thickens or thins text to improve recognition of characters.
- Border Removal or Addition: Aids in the efficient execution of text extraction.
- Skew and Orientation Correction: Aligns image and text properly for extraction. Skewed images may greatly reduce text extraction accuracy.
- Region of Interest (ROI) Identification: Detects regions of text through bounding boxes, according to specifications, refining target OCR zones.

3.2.6.2 Source I (b) Pre-processing

The data must be pre-processed for two primary reasons, these being anonymisation (blinding) and preparation for algorithmic training (AI model development).

Blinding

Removal and Abstraction of personal identifiers within information, ensuring that the data utilised achieves complete anonymity, enabling the utilisation of the dataset in an ethical and legally compliant manner.

Cleaning, Standardisation, and Resolution of Data Gaps

The dataset is cleaned to ensure that the data is meaningful and usable i.e. that it shall not produce unnecessary distortion in the model's results or functionality. Column data is standardised and manually checked for inconsistencies, this being feasible due to the relatively small sample size. Data gaps in less critical columns were filled with placeholders, allowing the records to be considered without distorting other columns, while gaps in critical columns, such as target columns were resolved with omission of record, ensuring all data utilised provides relevant, meaningful impact within the produced model.

3.2.6.3 Source II Pre-processing

The datasets shall be pre-processed in a similar manner as the Source I (b) dataset, highlighted at section 3.2.6.2 above.

3.3 System Architecture Design

3.3.1 Introduction

The system shall automate the importation, enhancement, and semantic analysis of HR case-related scanned documents and leverage AI to enable the automated processing of HR cases and the implementation of decision support to improve efficiency in the HR case processing domain, with the overall goal of automating the work procedures involved. The solution shall leverage a computer vision and NLP pipeline to enable downstream decision support and automated processing of formal correspondences. The architecture shall remain modular, combining image processing via OpenCV, OCR via Tesseract, and sentiment analysis using a basic aspect-based lexicon approach implemented in pure Python. This setup shall allow the extraction of structured sentiment insights from unstructured scanned text. An AI model shall additionally be implemented for the processing of digitised cases.

3.3.2 Planned Functional Modules

3.3.2.1 Image Importation (Input) and Display (cv2, matplotlib)

The module shall begin by importing scanned image files using the cv2.imread() function from the OpenCV module. For visualisation and inspection, a custom display() function shall be implemented using matplotlib.pyplot() with predefined resizing to allow for visual inspection of images and debugging of image processing.

3.3.2.2 Image Preprocessing (cv2)

The preprocessing stage shall consist of multiple image transformations aimed at enhancing downstream OCR output:

a) Grayscale Conversion: To convert images into grayscale, simplifying further binary processing.
b) Noise Removal: To remove image noise and enhance character contours.
c) Font Manipulation: Optional steps using dilation or erosion shall allow for font thickening or thinning to optimise character boundaries for recognition.
d) Inversion: To ensure optimal performance and compatibility with varied OCR engines.
e) Skew Correction: To align documents horizontally. A necessary step to optimise character line detection in text by ensuring proper document text orientation.
f) Border Removal and Cropping: To isolate content from document borders which may potentially distort or reduce accuracy achieved by OCR engine text extraction.
g) Thresholding: To binarize the image for further processing.
h) Dilation: To connect text characters into coherent line structures.
i) Region of Interest Detection: To detect and isolate text blocks for targeted OCR text extraction.
j) Bounding Box Detection: To define the region of interest (ROI) for OCR extraction.

3.3.2.3 Optical Character Recognition (pytesseract, NLTK)

The extracted ROIs shall be passed to pytesseract., which interfaces with the installed Tesseract engine. The OCR output shall be returned as raw text and stored in memory for processing. OCR quality shall be evaluated based on percentage accuracy of character extraction.

3.3.2.4 Sentiment Analysis (RegEx, NLTK)

This module shall implement a rule-based SA approach using a customised lexicon. The sentiment analysis module shall consist of:

a) Lexicon Design: Featuring two curated dictionaries (for positive and negative sentiment terms) that shall contain manually selected domain specific terms assigned appropriate weights.
b) Sentence Scoring: Extracted sentences shall be analysed through text matching and token parsing using Pythons re(Regular Expressions) module. Scores shall be

accumulated for each aspect identified and shall be used to produce a final overall sentiment score.

3.3.2.5 Database Connectivity (mysql.connector)

The system shall include a preconfigured connection to a local MySQL database planned to run for development purposes using phpMyAdmin with Apache. The database connection is to allow the passing of extracted key data points to a structured data store, which allows referential posterity and further processing, analysis or otherwise utilization as necessary.

3.3.2.6 AI Model Integration

Varied models are considered for implementation as an alternative to the Sentiment Analysis based approach to case processing and decision support. Varied algorithmic modelling techniques shall be assessed in order to identify the most viable model approach. Planned for testing are:

- XG Boost Model;
- Logistic Regression Model; and
- RandomForest Tree Model.

3.4 Prototyping

A Flowchart of the Proposed Solution is shown below:

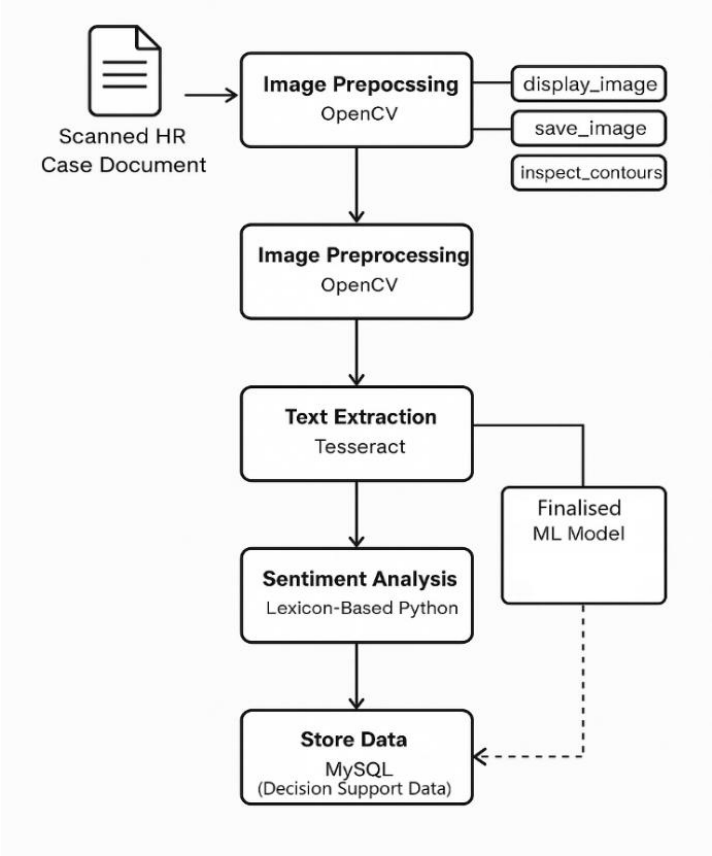


Figure 3. 3 Solution Flowchart

3.5 Evaluation Strategy

3.5.1 System Performance:

- Achievement status of outlined objectives of the research;
- Application Cycle Runtimes; and
- System Compute Requirements.

3.5.2 OCR Accuracy:

- Extracted Text Percentage Accuracy.

3.5.3 AI Models:

To evaluate the performance of the models developed, the following metrics shall be utilised:

- F1 and/or F2 Measure;
- PR Curve and/ or AUC;
- Confusion Matrix: Precision, Recall, FPR, TNR, FNR; and
- Accuracy.

3.6 Tools and Technologies

Summary Table of Tools and Uses

All Modules are developed within Geanny and Pycharm Community Edition utilizing the python programming language. A summary table of technologies utilised is indicated below:

Table 3. 2 Summary tools and Technologies

Module	**Technology**	**Purpose**
1. Image Importation	OpenCV, matplotlib	Import scanned HR case-related image files for initial processing and visualization.
2. Image Preprocessing	OpenCV	Enhance image quality by converting to grayscale, removing noise, manipulating fonts, and inverting the image for better OCR compatibility.

Module	Technology	Purpose
3. Optical Character Recognition (OCR)	Tesseract OCR Engine (pytesseract)	Convert detected regions of interest into raw text using Tesseract OCR for further analysis.
4. Sentiment Analysis	Custom Lexicon, RegEx, NLTK	Analyze sentiment in extracted text based on predefined positive/negative HR-specific terms and string matching to compute sentiment scores for key aspects.
5. Database Connectivity	MySQL (mysql.connector, phpMyAdmin, Apache)	Store processed output, sentiment values, or metadata in a local MySQL database for persistence and future analysis.
6. AI Model Integration	Pandas, Sklearn, Imblearn, matplotlib, joblib.	Prepare models through training and testing and evaluation through produced metrics. Final model selected saved for future use within project

3.7 Ethical Concerns related to Research

3.7.1 Introduction

The previous section detailed the data collection methods employed for the development of this thesis and project. This section explores the ethical concerns around the implementation and development project including employee displacement, system bias and data confidentiality concerns.

3.7.2 HR Professional Displacement and redundancy concerns

The question of whether something should be automated if it can be automated borders philosophical and ethical ground. On one hand the automation of a process could greatly improve provision of a service or completion of a work task, while on the other hand it may lead to the redundancy of previously valid and necessary jobs and of the individuals who perform them within those particular roles. With this preface established, it is understood that while automating the processing of HR case documents and their digitisation, care must be taken to ensure that the potential for displacement of vital personnel and the fears thereof are managed.

Within the context of this project HR professionals are regarded with importance, with the developed solution not to act in place of the human professional, but rather to augmentatively assist in the execution of tasks by incorporating its use into already established workflows. HR professional involvement is necessary in the successful implementation of the solution.

3.7.3 System Bias and Handling of Nuanced Situations

It must be stated that this research involves the automation of tasks traditionally and with reason performed by Human HR Professionals. This due to the sensitive nature of case files. While a system is able to provide a recommended or predicted solution to a problem in a domain where rules are clearly established, underlying concerns do exist as HR cases may on occasion present highly nuanced, multi-faceted information, which may require a more sensitive consideration approach. In cases such as those Human professionals are a better fit that the usage of strict rule-based logic, which may not cater to more subtle case details.

Within the context of this project, HR professional involvement shall be an integral part of the established workflow. Allowing for finer case details to be considered before the rendering of judgements.

With regards to avenues for future improvements to incorporate or bolster solution decision making with regards to finer nuanced aspects, fuzzy logic may be incorporated to factor in degrees of membership of aspects and sentiments. Additionally, an alternative avenue of improvement may involve the use of machine learning models trained off anonymised real world case data, with the removal or muting of attributes that may contribute to case processing bias.

3.7.4 Data Protection and Confidentiality Concerns

Due to the nature of human resource cases containing personal and in many cases, sensitive personal information, applicable legislation and regulations in the locale in which the system is implemented for use must be taken into account. With the current locale of the implementation being the Government Republic of Zambia, the Laws of Zambia must be considered.

The Data Protection Act, 2021 of the Laws of Zambia, establishes regulations on the storage and processing of personal information, including legislation on the valid terms of storage and usage of personal information. The Electronic Government Act, No. 41 of 2021

additionally contains regulations on the processing of information as well as the movement of data within and across public bodies.

As such, the above-mentioned legislative documents shall be considered during project development and implementation. Best practice of systems development shall be adhered to, with strict emphasis on the systems development and testing using only a small set of real-world case documents. The considered datasets shall be generated or anonymised to ensure the protection of personal information.

3.8 Chapter Summary

The chapter focused on several aspects related to the project's implementation. This chapter tackled the detailing of the selected research design, research methods selected, the data upon which the research was based and the data with which the project shall be implemented. This chapter also details a discussion on the ethical concerns around the development of the project and the data that shall be part of its pipeline. The next chapter seeks to explore the technical aspects of the project with regard to data, experimentation and its implementation.

CHAPTER 4 – DATA, EXPERIMENTS, AND IMPLEMENTATION

4.1 Appropriate modelling in relation to project

4.1.1 Introduction

The modelling approach selected for the development of this project is Evolutionary Prototyping. This model's use is intended to create a scalable, replicable and efficient framework for processing HR case documents. This process involves the integration of advanced image processing techniques, utilizing Optical Character Recognition, regular expressions (RegEx), Sentiment Analysis, and a structured data storage mechanism. Additionally, the use of an AI model is employed and explored as an evolutionary alternative solution to the case processing automation and decision support earlier implemented through the use of Sentiment Analysis.

The goal of this development is to automate HR case processing, digitise paper-based documents, extract key information, and store the extracted and produced data in a structured and accessible format, facilitating improved workflows in HR case management. This section explores the appropriateness of the selected model, technologies involved, and their integration into the project.

4.1.2 Modelling Techniques and Technologies

The model selected for use in the development of this work is the Evolutionary Prototyping model. It allows the solution to iteratively, and with continually improving functionally incorporate several advanced technologies, each addressing specific challenges in document digitisation and processing.

4.1.3 Justification for use of Evolutionary Prototyping Model

The selected development model offers multiple advantages in alignment with the DSR Methodology, including:

- Rapid Modular Development: The model allows for rapid development of a solution in a modular manner. Producing singular functional and testable modules that can be expanded upon in a prompt manner. This is beneficial as individual components can be implemented and reviewed, allowing for rapid resolution of issues and improvement of features. This overall, allowing the solution to evolve in a fully functional manner, with opportunities for revision and refinement.

- Iterative Development: Iterative development is essential to the continuous refinement of a solution. The solution requiring development involves the integration of several complex technologies in a cohesive and evolving manner. Practically allowing for the evolution of modules developed as well as approaches taken to the implementation of problem-solving solutions, such as with the initial development of a Sentiment Analysis Module, which may be alternated with an AI model for classification developed later, thus continually improving the solution and its implemented technologies. Iterative development as well allows for the efficient resolution of discovered issues.

- Frequent Evaluation: The model allows for constant evaluation of developed modules, allowing for the final solution to be refined in a quicker manner, with the main priority being the production of a functional prototype of the whole solution or modules of the solution.

4.2 Development

4.2.1 Introduction

The developed solution was built sequentially using a modular development style. These developed models were then integrated to assemble the full functional proposed framework for automated, efficient human resource case processing. The framework constitutes multiple developed modules built with Python programming language with clearly defined purposes, each. The framework integrates image pre-processing, computer vision, natural language processing, and machine learning to automate and streamline the process of HR case document processing in an efficient manner. The modular architecture implemented promotes flexibility and robustness.

4.2.2 Developed Framework Modules

The components (modules) of the framework developed are:

4.2.2.1 Framework Diagram

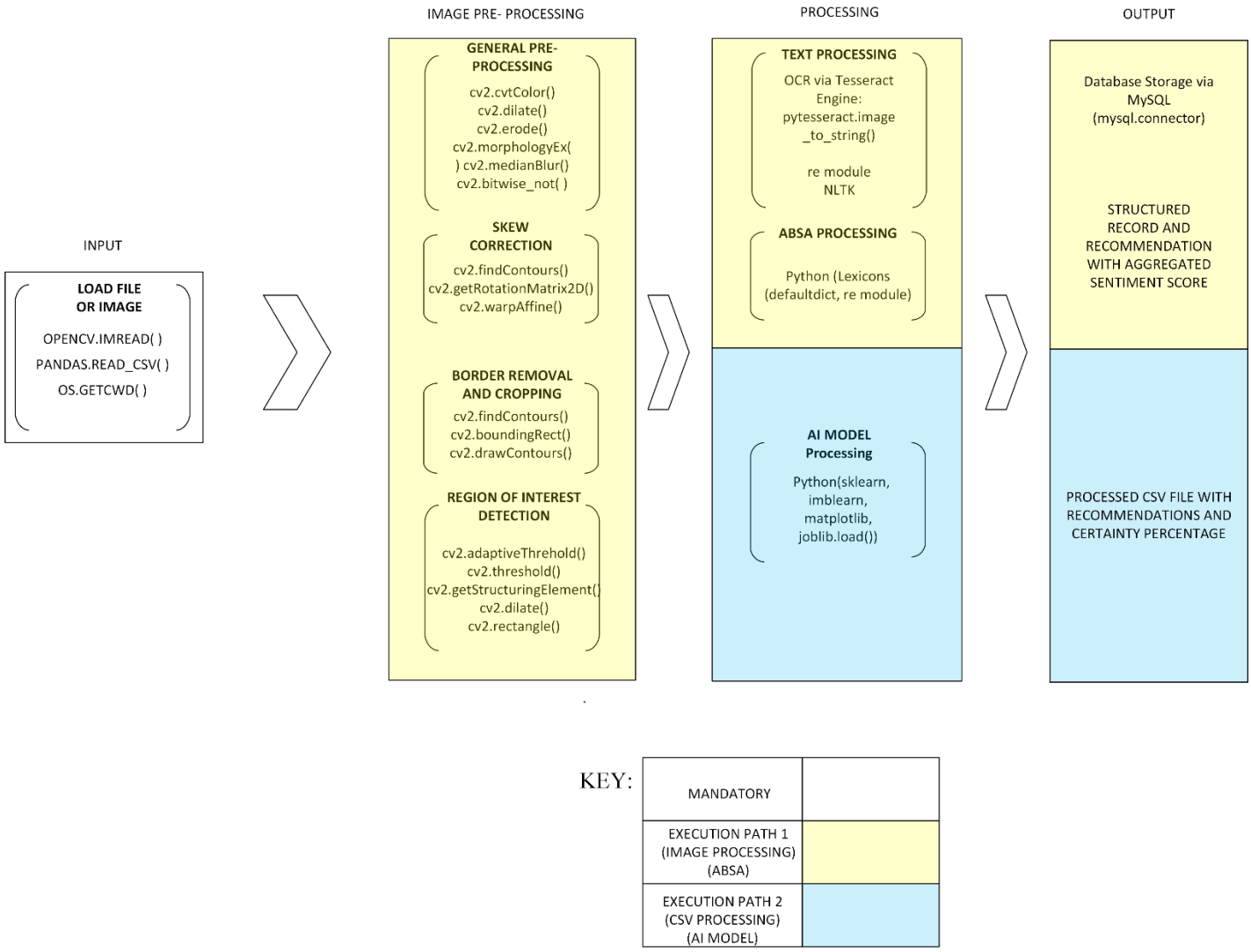


Figure 4. 1 Framework Module Diagram

4.2.2.2 Image importation (Input) module

This module utilizes OpenCV's cv2.imread() function to select scanned image files for input and subsequent processing. A custom display function was prepared to allow for visual review of the selected image files, from initial appearance through all pre-processing steps in preparation for use by subsequent modules.

4.2.2.3 Image Pre-processing Module

This module features multiple mandatory and optional steps, largely utilizing functions of the OpenCV python module for the processing of images. These steps include:

- Grayscale Conversion: Using cv2.cvtColor() function to simplify image data in preparation for subsequent binarization using Otsu's method [97].
- Noise Removal: Utilizing the cv2.dilate(), cv2.erode() and cv2.medianBlur functions to mitigate image noise.
- Font Modification: An optional operation using the cv2.dilate() and cv2.erode() functions to thicken or thin font as necessary for the improved detection of characters when OCR is performed.
- Image Inversion: Using the cv2.bitwise_not() function to ensure OCR compatibility of images.
- Skew Angle Correction: An optional step using a custom developed deskew() function applied to correct orientation of images and image text in order to improve OCR accuracy.
- Border Removal: An optional step using multiple cv2 module functions to remove any present borders which may distort the results of a performed OCR scan.

4.2.2.4 Region of Interest (ROI) Detection Module

This module is used to obtain regions of interest, i.e. targeted locations with extractable text identified on images for processing. This module is executed with three main functions and utilises the pre-processed image prepared from the earlier Grayscale Conversion function. The following steps are executed within this module:

- Image Thresholding: Gaussian image theresholding and utilization of Otsu's method [97] are utilised.
- Image Dilation: The image is horizontally dilated using the cv2 function cv2.getStructuringElement() and cv2.dilate() functions to connect text into blurred together lines, producing a silhouette of clustered text region.
- Bounding Box Drawing: Contours are drawn around the identified blurry text boxes using cv2.rectabgle() which define regions of interest for OCR processing at a later stage.

4.2.2.5 OCR Module

ROIs earlier obtained are passed into the pytesseract.image_to_string() function to extract the raw text from the specified regions. The OCR module then leverages the python re (RegEx) module for post-OCR text cleanup and parsing. Through regular expressions, the text is tokenised in a custom manner and filtered to identify case/domain specific patterns and data

points, enhancing the precision of the aspect identification and sentiment mapping in a later stage of the text processing.

```
# Regex Function
patterns = [
    (r'M(r\.|rs\.|s\.)', "Title"),  # Title
    #   (r'(M(r\.|rs\.|s\.)\s)(\w+\s\w+)', "First and Last Name"),  # First and Last Name
    #   (r'\(TS/\d+\)', "TS Number"),  # TS Number
    #   (r'\(NRC\/\d{6}\/\d{2}\/1\)', "NRC Number"),  # NRC Number
    (r'(?i)(Head\s+Teacher|Deputy\s+Head\s+Teacher|Subject\s+Teacher|Class\s+Teacher|Head\s+of\s+Department\s+-\s+[^\s(]+(?:\s+[^\s(]+)*)',
     "Job Title"),  # Job Title
    (r'\({1}[A-Z]{1}\){1}', "Salary Scale"),  # Salary Scale
    (r'(\b\w+\b)\s+(\b\w+\b)\s+(\b\w+\b)\s+(?=School\b|school\b)', "Name of School"),  # Name of School
    (r'(\b\w+\b)\s+(?=[Dd]istrict)', "District"),  # District
    (r'(\b\w+\b)\s+(?=[Pp]rovince)', "Province")  # Province
]
name_pattern = r'M(r\.|rs\.|s\.)\s\w+\s\w+'
TS_Number_pattern = r'\(TS\.\s\d+\)'
NRC_Number_pattern = r'\(NRC\/\d{6}\/\d{2}\/1\)'

```

Figure 4. 2 Regex Matching Patterns Code Snippet

4.2.2.6 Sentiment Analysis Module

The sub-module, integrated into the OCR Module employs a rule-based Sentiment Analysis pipeline utilising a custom developed lexicon for the mapping of identified text expressions to sentiments, positive or negative, which may predict and make recommendations on the processing of an HR case. This approach was intended to be evolutionarily replaced with a theoretically more efficient developed AI classification model.

4.2.2.7 Database Storage Module

A module was developed for the integration of a database solution using mySQL within a local server environment. The database connection is made possile using the python mysql.connector() function. The local server environment is set up using Xampp, Apache and PhpMyAdmin. The data storage pipeline allows for derived case aspects (obtained through OCR, RegEx and SA functionality) to be stored in a structured manner for all relevant utility and referential posterity.

#	Name	Type	Collation	Attributes	Null	Default	Comments	Extra
1	ID	int(11)			No	None		AUTO_INCREMENT
2	file_code	varchar(255)	latin1_swedish_ci		Yes	NULL		
3	doc_author	varchar(255)	latin1_swedish_ci		Yes	NULL		
4	doc_author_pos	varchar(255)	latin1_swedish_ci		Yes	NULL		
5	doc_support_status	varchar(50)	latin1_swedish_ci		Yes	NULL		
6	case_type	varchar(255)	latin1_swedish_ci		Yes	NULL		
7	case_name_1	varchar(255)	latin1_swedish_ci		Yes	NULL		
8	case_name_2	varchar(255)	latin1_swedish_ci		Yes	NULL		
9	case_ts_no_1	varchar(255)	latin1_swedish_ci		Yes	NULL		
10	case_ts_no_2	varchar(255)	latin1_swedish_ci		Yes	NULL		
11	case_nrc_no_1	varchar(255)	latin1_swedish_ci		Yes	NULL		
12	case_nrc_no_2	varchar(255)	latin1_swedish_ci		Yes	NULL		
13	case_job_title_1	varchar(255)	latin1_swedish_ci		Yes	NULL		
14	case_job_title_2	varchar(255)	latin1_swedish_ci		Yes	NULL		
15	case_salary_scale_1	varchar(255)	latin1_swedish_ci		Yes	NULL		
16	case_salary_scale_2	varchar(255)	latin1_swedish_ci		Yes	NULL		
17	case_school_1	varchar(255)	latin1_swedish_ci		Yes	NULL		
18	case_school_2	varchar(255)	latin1_swedish_ci		Yes	NULL		
19	case_district_1	varchar(255)	latin1_swedish_ci		Yes	NULL		
20	case_district_2	varchar(255)	latin1_swedish_ci		Yes	NULL		
21	case_province_1	varchar(255)	latin1_swedish_ci		Yes	NULL		
22	case_province_2	varchar(255)	latin1_swedish_ci		Yes	NULL		
23	aa_or_tfer_type	varchar(50)	latin1_swedish_ci		Yes	NULL		
24	case_aa_period	varchar(50)	latin1_swedish_ci		Yes	NULL		
25	case_date	date			Yes	NULL		
26	Sentiment	varchar(255)	latin1_swedish_ci		Yes	NULL		
27	Paragraph	text	latin1_swedish_ci		Yes	NULL		

Check all *With selected:* Browse Change Drop Primary Unique Index

Figure 4. 3 Prototype Database Schema and Datatypes

Table 4. 1 Implemented modules and technologies

Module	Technology	Purpose
1. Image and File Importation	OpenCV (cv2.imread()) OS os.getcwd() Pandas pandas.read_csv()	Import scanned HR case-related image files for initial processing and visualization.
2. General Image Preprocessing	OpenCV (cv2.cvtColor(), cv2.dilate(), cv2.erode(), cv2.morphologyEx(), cv2.medianBlur(), cv2.bitwise_not())	Enhance image quality by converting to grayscale, removing noise, manipulating fonts, and inverting the image for better OCR compatibility.
3. Skew Correction	OpenCV (cv2.findContours(), cv2.getRotationMatrix2D(), cv2.warpAffine())	Correct document skew by detecting the angle and rotating the image for alignment, optimizing text line detection.
4. Border Removal and Cropping	OpenCV (cv2.findContours(), cv2.boundingRect(), cv2.drawContours())	Detect and remove borders, such as stamps and folds, to isolate content for OCR.
5. Region of Interest Detection	OpenCV (cv2.adaptiveThreshold(), cv2.threshold(), cv2.getStructuringElement(), cv2.dilate(), cv2.rectangle())	Detect textual blocks by thresholding, dilating, and drawing bounding boxes around areas of interest for OCR extraction.
6. Optical Character Recognition (OCR)	Tesseract via pytesseract.image_to_string()	Convert detected regions of interest into raw text using Tesseract OCR for further analysis.
7. Aspect-Based Sentiment Analysis (ABSA)	Python (Lexicons (defaultdict, re module)	Analyse sentiment in extracted text based on predefined positive/negative HR-specific terms and string matching to compute sentiment scores for key aspects.
8. Artificial Intelligence Model	Python(sklearn, imblearn, matplotlib, joblib)	Analyse features of human resource case to determine recommended course of action with degree of certainty of recommendation.
8. Database Connectivity	MySQL (mysql.connector)	Store processed output, sentiment values, metadata in a local MySQL database for persistence and future analysis.
9. Debugging & Modular Testing	OpenCV (cv2.imshow(), cv2.imwrite()), Custom Wrapper Functions	Visualize and save intermediary processing steps for debugging and tuning, providing transparency in the preprocessing pipeline.

4.2.3 AI Classification Module and Model Training

4.2.3.1 Baseline (Control) Logistic Model – Trained and Tested on Repository Dataset

To establish a performance baseline for model performance metrics, a Logistic Regression with 80:20 sampling respectively for training and testing. The baseline model was trained and tested using a separate repository sourced employee promotion dataset [96]. Subsequent Modules were trained using the Primary Institution-Provided Dataset. A baseline logistic regression model was as well utilised with the Primary Institution-Provided Dataset.

4.2.3.2 Logistic Model – Trained and Tested on Institution-Provisioned Dataset

To gauge the performance of the Logistic Regression Model when tested with the functional dataset, the baseline logistic regression model was then tested and trained on the institution provisioned dataset- a smaller sample size with a significantly increased class imbalance due to the small minority class sample size. This variation in sample set size expectedly resulted in performance statistic reductions across multiple statistics in comparison to the control model.

To accommodate the imbalance of the dataset and with the recognition that that the metric of accuracy metric may be less informative within the context of the provisioned dataset, the F2-Measure was included as a key evaluation metric. The F2-Score was selected due to its emphasis on recall, making it particularly suitable for the assessment of the model's ability to the omission of minority class cases.

4.2.3.3 Logistic Regression Model with SMOTE (Synthetic Minority Oversampling Technique) - Trained and Tested on Institution-Provisioned Dataset

To improve the overall performance of the Logistic Regression Model, a SMOTE process was integrated as an attempt at addressing the imbalance within the dataset's minority class. The usage of the data augmentative technique generally improved the performance of the model across all metrics but PR AUC, majority class Precision and majority class Recall, which remained static.

4.2.3.4 Random Forest Classification Model - Trained and Tested on Institution-Provisioned Dataset

A random forest classification model was developed with the purpose of deriving insights into the achievable performance when prioritising Majority Class Precision and Recall as well as to examine feature importance. As anticipated, the model's performance metrics were

heavily skewed in favour of the majority class– predicting Promoted employees exceptionally well, whilst offering comparatively limited recall for the minority class.

The Random Forest model achieved higher minority class precision, indicating an improved ability to correctly identify True Positive (TP) minority cases when predictions were made.

4.2.3.5 Random Forest Classification Model with SMOTE+ENN (Edited Nearest Neighbours) - Trained and Tested on Institution-Provisioned Dataset

The Random Forest model was augmented with the addition of both a SMOTE process, for the oversampling the minority class, and the ENN technique to maintain data quality by removing noisy and borderline data samples. These techniques were implemented with the goal of achieving greater model performance.

The model achieved a moderate improvement in minority class recall; however, this came at the cost of a notable reduction in minority class precision. In addition, slight improvements in majority class precision and F1-Score were recorded over the base Random Forest Model. PR AUC remained unchanged, and all other recorded metrics declined, with minority class precision notably reducing to the lowest recorded value among all tested models (0.09).

4.2.3.6 Random Forest Classification Model with Balanced Bagging - Trained and Tested on Institution-Provisioned Dataset

A standard balanced bagging approach with majority class under-sampling utilizing a group of 10 learners was implemented as an alternative augmentation to the random forest model as well as a general approach to achieve higher performance metrics than the Random Forest with SMOTEENN model.

The standard balanced bagging approach notably achieved the highest recorded minority class recall (0.89), and additionally yielding slight increases in PR AUC Score, majority class Precision, and minority class F1-Score. All other metrics observed indicated minor decreases.

4.2.3.7 Random Forest Classification Model with Balanced Bagging and SMOTEENN - Trained and Tested on Institution-Provisioned Dataset

The balanced bagging model was further augmented with the introduction of SMOTE and ENN in order to increase the performance of the model on the minority class through under-sampling of the majority class as well as sample cleaning through ENN and parameters set to

limit any data pruning to non-minority class items. The model was configured in additionally modified to utilise a group of 6 learners- identified as the optimal learner configuration through experimentation.

The augmented balanced bagging Random Forest model with SMOTE and ENN achieved moderate performance score improvements generally across the minority class metrics. The model achieved an F1 Accuracy score of 0.95, a weighted average F1-Score of 0.95 and a majority class F1-Score of 0.97- equalling the previously achieved highest recorded scores of the mentioned metrics, all achieved by the base Random Forest model. Additionally, the model achieved the highest recorded minority class Precision (0.31), F1-Score (0.36) and F1 Macro Average (0.67). In addition, the model achieved excellent majority class precision (0.98) and recall (0.96) scores as well as a high F2-Score of 0.97 and a PR AUC Score of 0.98. There was observed, however, a moderate reduction in minority class recall of roughly 50% in comparison to the standard Balanced Bagging Random Forest model.

The development of the AI models was iteratively executed with multiple models evaluated. The process ultimately concluded with the selection and integration of the final selected model as the trial AI Module component within the developed solution. The **Random Forest Classification Model with Balanced Bagging and SMOTEENN** was selected for integration in the AI Module, as it achieved relatively higher performance metrics with regards to minority class predictions, as well as fair majority class performance. The data set imbalance was considered in the model selection process, as a model that best fits the practical needs of real-world scenarios in the operating environment within which the system is intended for use shall measure most ideal.

At the conclusion of the model's development the RandomForest Classifier Model was selected as most optimal of the evaluated models. The reasoning for the selection is that the model produced better metric scores than both Logistic Regression Models and possessed excellent scores on the metrics of F2 Measure and PR AUC score.

4.2.3.8 AI Model Selected for AI Module Integration

The **Random Forest Classification Model with Balanced Bagging and SMOTEENN** was selected for integration in the AI Module. The module was developed and configured as shown in the table below:

Table 4. 2 AI Model Development and Configuration

Development Task	Implementation
1. Data Preparation and Feature Engineering	- Raw HR data (SP-.csv) loaded via Pandas. - Irrelevant fields removed (e.g., NAME, DATE OF BIRTH). - AGE used as a numeric substitute. - Categorical features (SEX, DISTRICT, PROVINCE, POSITION, QUALIFICATION) one-hot encoded via SimpleImputer + OneHotEncoder in a pipeline.
2. Class Imbalance Handling	- SMOTEENN applied: combines SMOTE (oversampling the minority class) with Edited Nearest Neighbours (ENN) cleaning to remove noisy majority class samples only. - ENN parameter: sampling_strategy='not minority'.
3. Model Selection and Training	- BalancedBagging Classifier (6 Learners) with Random Forest base estimators. - Preprocessing and resampling integrated via ImbPipeline. - Stratified split: 80% training, 20% testing using train_test_split(stratify=y).
4. Model Evaluation and Metrics	- Precision-Recall Curve - PR AUC - F1-Score - F2-Score - Confusion Matrix - Classification Report
5. Model Output and Saving	- Trained pipeline saved using joblib.dump() for deployment as balancedbagging_smote_randomforest_model.pkl.
6. Data Pipeline and Integration	- Full modular pipeline includes: preprocessing → SMOTEENN resampling → Balanced Bagging Classifier. - Easily extendable for deployment or retraining workflows.

4.2.3.9 Other model(s) evaluated:

- XGBoost (XGBClassifier) Model – Scored perfectly in terms of metrics, this perfection was then identified as a result of overfitting, rendering the model ineffective. No further experimentation was carried out with this model.

4.3 Testing

Introduction

The modules were tested through the use of varied metrics and visual inspections. Screenshots of the testing process are shown, however, screenshots may be censored or omitted at certain stages to protect sensitive information.

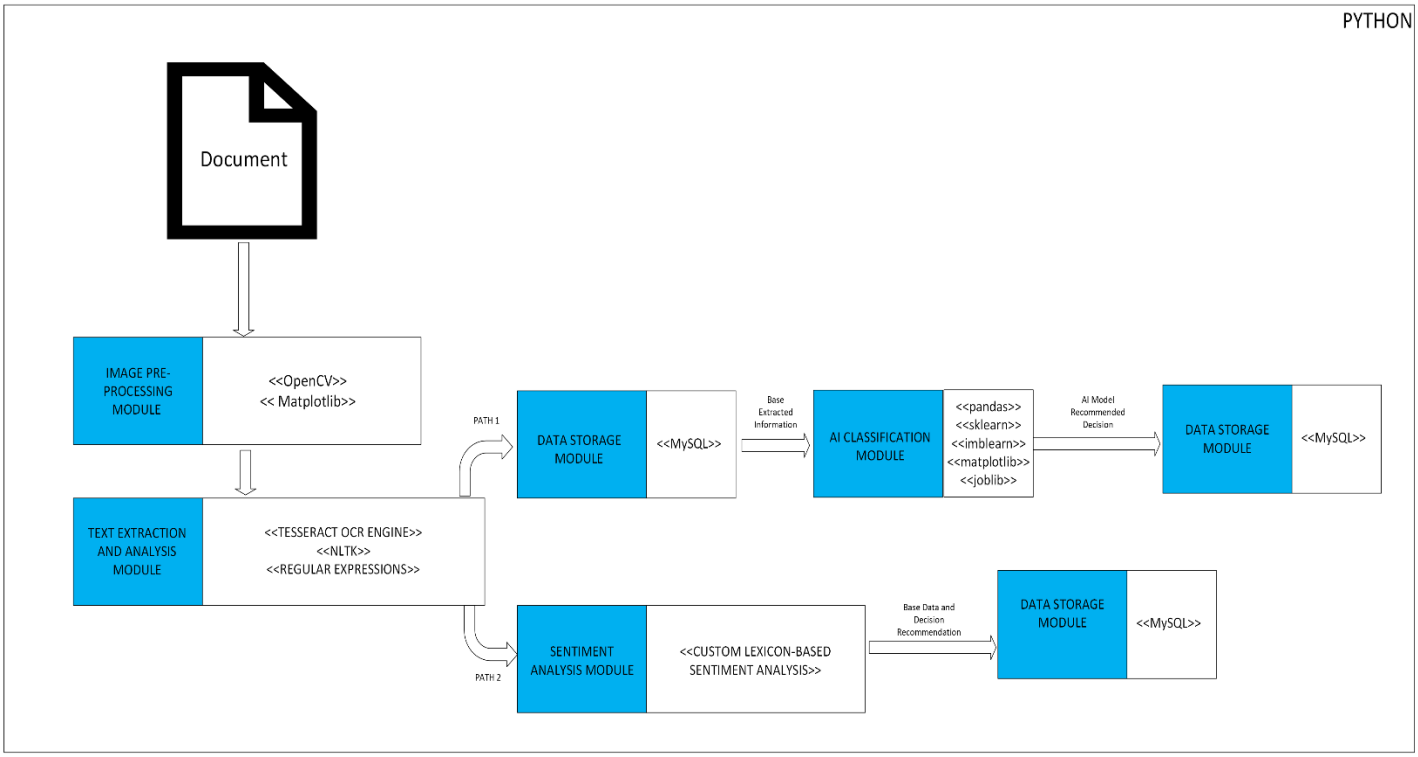


Figure 4. 4 Module Data Pipeline

4.3.1 Module Testing

In order to test the functionality of the Modules involved in the pipeline, testing began with the OCR Module to gauge whether information was accurately and relevantly extracted with key data points isolated using RegEx, outputs of image processing iterations. These outputs were displayed through the compiler console to enable visual analysis of the results. The following screenshots show the pipeline from the display of extracted data to the insertion and storage of data in the integrated database solution.

```
--- Identified Result for ite_id: 9 ---
ite_id: 9
------------------------------
ABSA_result: ["{'positive': {'average_score': 0.85, 'count': 1}, 'neutral': {'average_score': 0.0, 'count': 1}, 'overall': {'average_score': 0.42, 'sentiment': 'positive'}}"]
------------------------------
Full Name: [redacted]
------------------------------
TS Number: [redacted]
------------------------------
case_type: ['Acting Appointment']
------------------------------
effect_date: [redacted]
------------------------------
prospect: ['Substantive Promotion']
------------------------------
Acting Position: Head of Department - Natural Sciences
------------------------------
Title: [redacted]
------------------------------
NRC Number: [redacted]
------------------------------
Job Title: Head of Department - Natural Sciences
------------------------------
Salary Scale: (J)
------------------------------
Name of School: [redacted]
------------------------------
District: Kabompo
------------------------------
Province: Western
------------------------------
Paragraph: [redacted]
------------------------------
Job Title_2: Head Teacher
------------------------------
Salary Scale_2: (K)
------------------------------
Name of School_2: [redacted]
------------------------------
Effect Date: Duty Assumption Date
------------------------------
<------------>

--- Identified Result for ite_id: 10 ---
```

Figure 4. 5 Extracted Data Console Output

```
File Code: [redacted], Support: Supported, HR Name: [redacted], HR Title: [redacted]
Data inserted successfully. ID: 68

Process finished with exit code 0
```

Figure 4. 6 Additional Case Batch Information

```
            case_job_title_1 = identified_result.get('Job Title', '')
            case_job_title_2 = identified_result.get('Acting Position', '')
            case_salary_scale_1 = identified_result.get('Salary Scale', '')
            case_salary_scale_2 = identified_result.get('Salary Scale_2', '')
            case_school_2 = identified_result.get('Transfer School Name', [''])[0]
            case_district_1 = identified_result.get('District', '')  #
            case_district_2 = identified_result.get('District_2', [''])[0]
            case_province_1 = identified_result.get('Province', '')  #
            case_province_2 = identified_result.get('Province_2', [''])[0]
            aa_or_tfer_type = identified_result.get('prospect', '')
            effect_date = identified_result.get('effect_date', '')
            Paragraph = identified_result.get('Paragraph', [''])[0]

            values = (
                file_code, doc_author, doc_author_pos, doc_support_status, case_type,
                case_name_1, case_name_2, case_ts_no_1, case_ts_no_2,
                case_nrc_no_1, case_nrc_no_2, case_job_title_1, case_job_title_2,
                case_salary_scale_1, case_salary_scale_2, case_school_1, case_school_2,
                case_district_1, case_district_2, case_province_1, case_province_2,
                aa_or_tfer_type, case_aa_period, case_date, Sentiment, Paragraph
            )
            values = tuple(', '.join(v) if isinstance(v, list) else v for v in values)

            cursor.execute(insert_query, values)

        # Commit the transaction
        conn.commit()

        print(f"Data inserted successfully. ID: {cursor.lastrowid}")

    except mysql.connector.Error as err:
        print(f"Error: {err}")

    finally:
        if conn.is_connected():
            cursor.close()
            conn.close()
```

Figure 4. 7 SQL Database Insertion Record Code Snippet

Figure 4. 8 Inserted Database Record

Figure 4. 9 Inserted Database Record Continuation

The results indicated by the visual outputs produced showcase the performance of the various integrated module frameworks, allowing for successful data extraction from scanned documents as well as their digitisation through storage in a structured database. Below are the test results of the AI Models built and evaluated for the purpose of providing a potentially superior alternative to the Sentiment Analysis Approach to prediction.

4.3.2 AI Model Evaluation

The dataset utilised for training was greatly skewed towards a dominant class. This was a factor that required consideration in the training of the module. Though the skewing imbalance did affect the finalised Model, it is believed to be a functional model, and the highest performing of all models evaluated in the development process.

```
[5 rows x 14 columns]
is_promoted
0    50140
1     4668
Name: count, dtype: int64
[[9991   63]
 [ 678  230]]
              precision    recall  f1-score   support

           0       0.94      0.99      0.96     10054
           1       0.78      0.25      0.38       908

    accuracy                           0.93     10962
   macro avg       0.86      0.62      0.67     10962
weighted avg       0.92      0.93      0.92     10962
```

Figure 4. 10 Metrics of Control Model - Logistic Regression Model – Trained and Tested on Repository Dataset

```
ACTION
PROMOTED        1268
NOT PROMOTED      44
Name: count, dtype: int64
F2 Score: 0.81
PR AUC: 0.99
Confusion Matrix:
 [[  7   2]
 [ 57 197]]
Classification Report:
              precision    recall  f1-score   support

NOT PROMOTED       0.11      0.78      0.19         9
    PROMOTED       0.99      0.78      0.87       254

    accuracy                           0.78       263
   macro avg       0.55      0.78      0.53       263
weighted avg       0.96      0.78      0.85       263
```

Figure 4. 11 Metrics of Logistic Regression Model – Trained and Tested on Provisioned Dataset

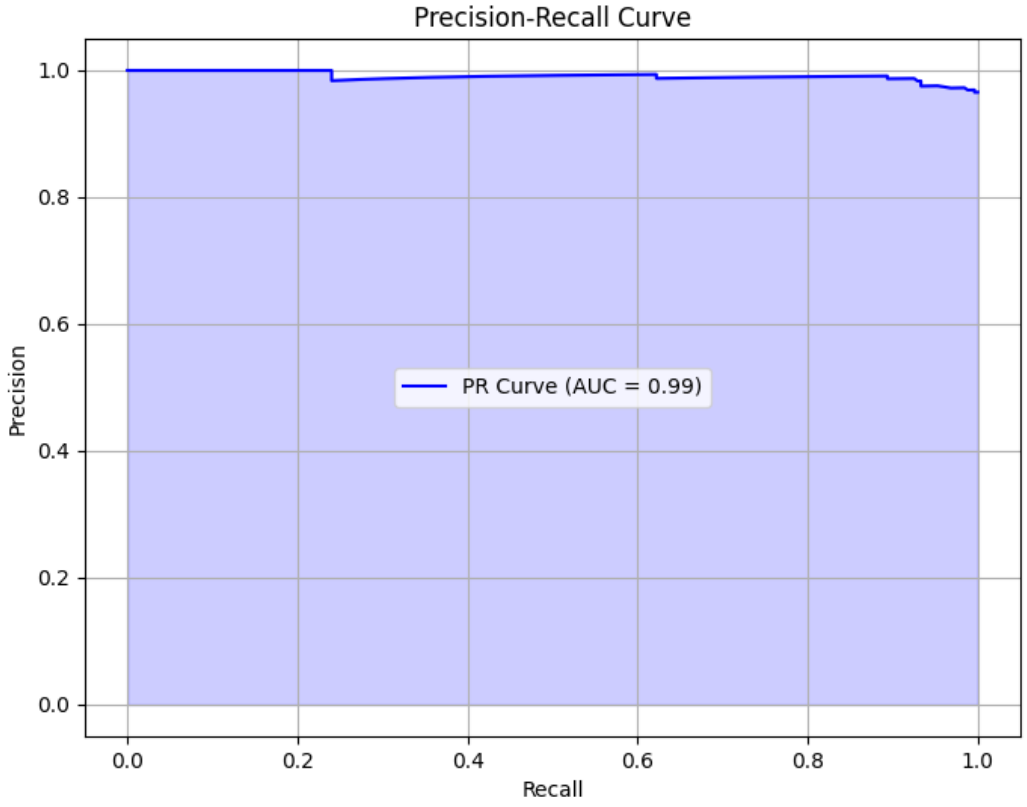


Figure 4. 12 PR Curve of Logistic Regression Model – Trained and Tested on Provisioned Dataset

```
ACTION
PROMOTED        1268
NOT PROMOTED      44
Name: count, dtype: int64
F2 Score: 0.84
PR AUC: 0.99
Confusion Matrix:
 [[  7   2]
 [ 48 206]]
Classification Report:
               precision    recall  f1-score   support

NOT PROMOTED       0.13      0.78      0.22         9
    PROMOTED       0.99      0.81      0.89       254

    accuracy                           0.81       263
   macro avg       0.56      0.79      0.56       263
weighted avg       0.96      0.81      0.87       263
```

Figure 4. 13 Metrics of Logistic Regression Model with SMOTE – Trained and Tested on Provisioned Dataset

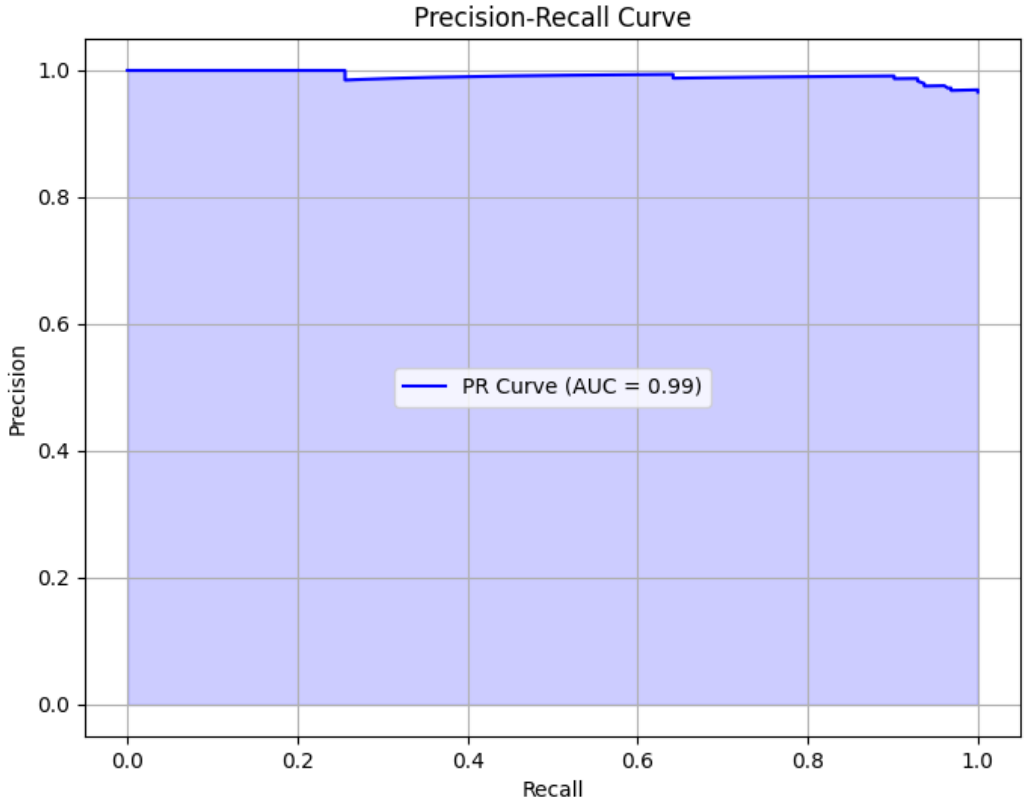


Figure 4. 14 PR Curve of Logistic Regression Model with SMOTE – Trained and Tested on Provisioned Dataset

```
ACTION
PROMOTED        1268
NOT PROMOTED      44
Name: count, dtype: int64
F2 Score: 0.98
PR AUC: 0.98
Confusion Matrix:
 [[  1   8]
 [  5 249]]
Classification Report:
              precision    recall  f1-score   support

NOT PROMOTED       0.17      0.11      0.13         9
    PROMOTED       0.97      0.98      0.97       254

    accuracy                           0.95       263
   macro avg       0.57      0.55      0.55       263
weighted avg       0.94      0.95      0.95       263
```

Figure 4. 15 Metrics of Random Forest Model with SMOTE – Trained and Tested on Provisioned Dataset

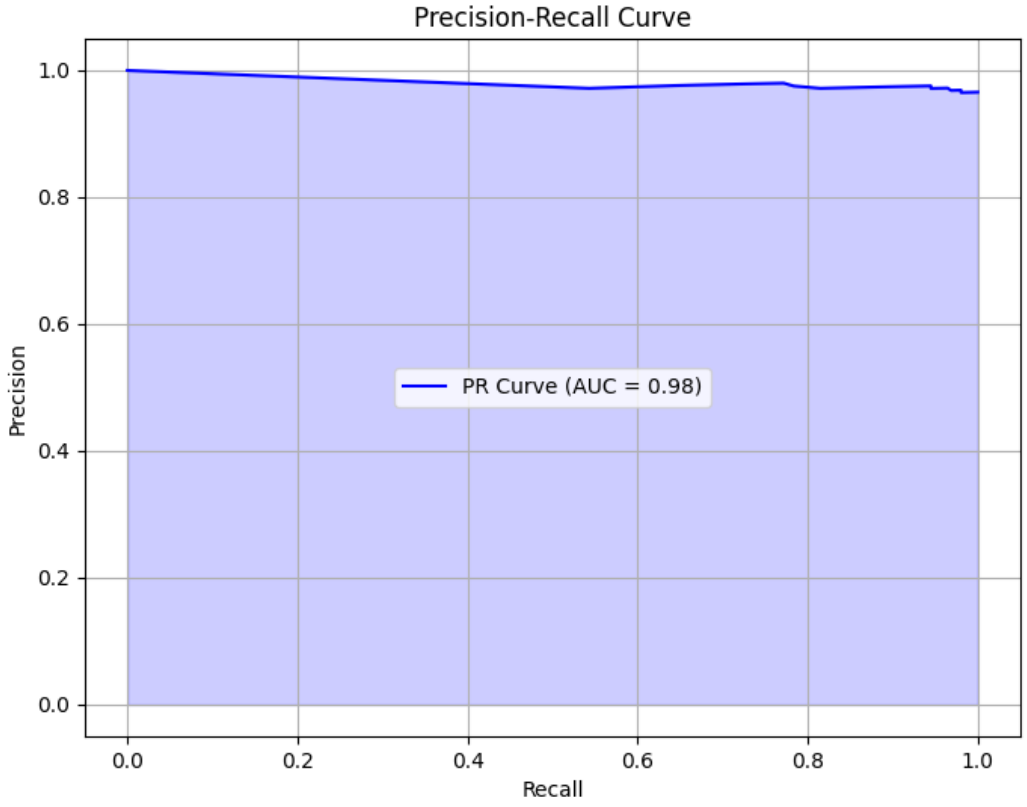


Figure 4. 16 PR Curve of Random Forest Model with SMOTE – Trained and Tested on Provisioned Dataset

```
ACTION
1    1268
0      44
Name: count, dtype: int64

F2 Score: 0.86
PR AUC: 0.98
Confusion Matrix:
 [[  4   5]
 [ 41 213]]
Classification Report:
              precision    recall  f1-score   support

           0       0.09      0.44      0.15         9
           1       0.98      0.84      0.90       254

    accuracy                           0.83       263
   macro avg       0.53      0.64      0.53       263
weighted avg       0.95      0.83      0.88       263
```

Figure 4. 17 Metrics of Random Forest Model with SMOTE-ENN – Trained and Tested on Provisioned Dataset

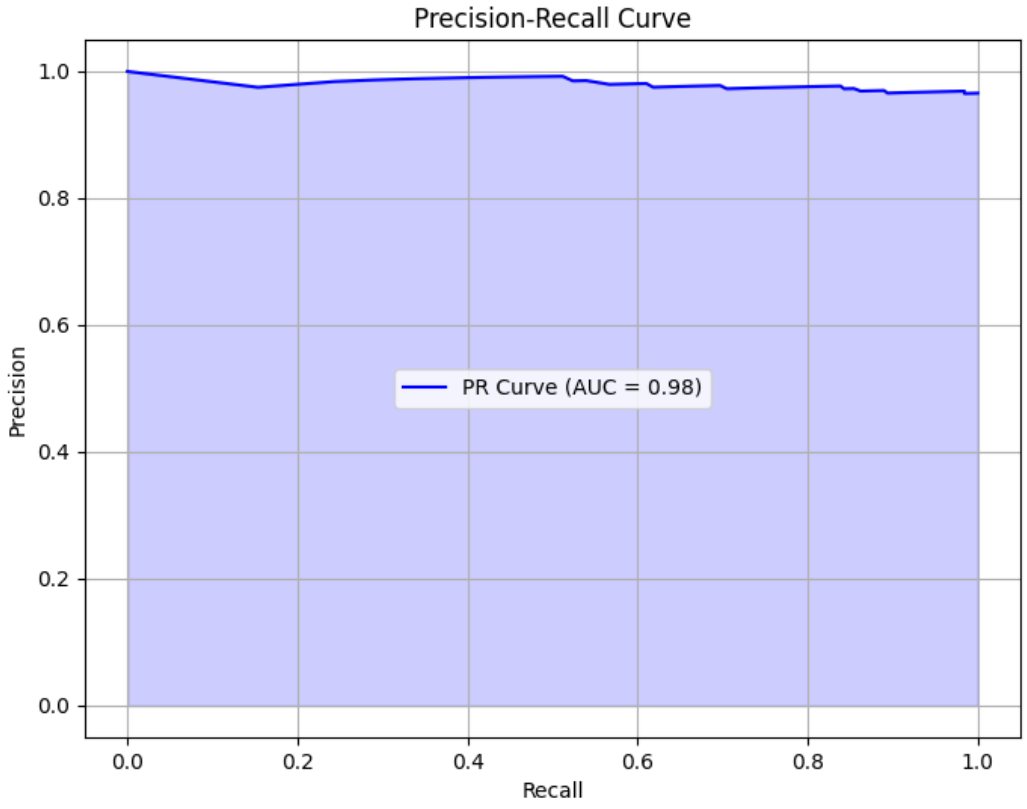


Figure 4. 18 PR Curve of Random Forest Model with SMOTE-ENN – Trained and Tested on Provisioned Dataset

```
ACTION
PROMOTED        1268
NOT PROMOTED      44
Name: count, dtype: int64
F2 Score: 0.78
PR AUC: 0.99
Confusion Matrix:
 [[  8   1]
 [ 66 188]]
Classification Report:
              precision    recall  f1-score   support

NOT PROMOTED       0.11      0.89      0.19         9
    PROMOTED       0.99      0.74      0.85       254

    accuracy                           0.75       263
   macro avg       0.55      0.81      0.52       263
weighted avg       0.96      0.75      0.83       263
```

Figure 4. 19 Metrics of Balanced Bagging Random Forest Model – Trained and Tested on Provisioned Dataset

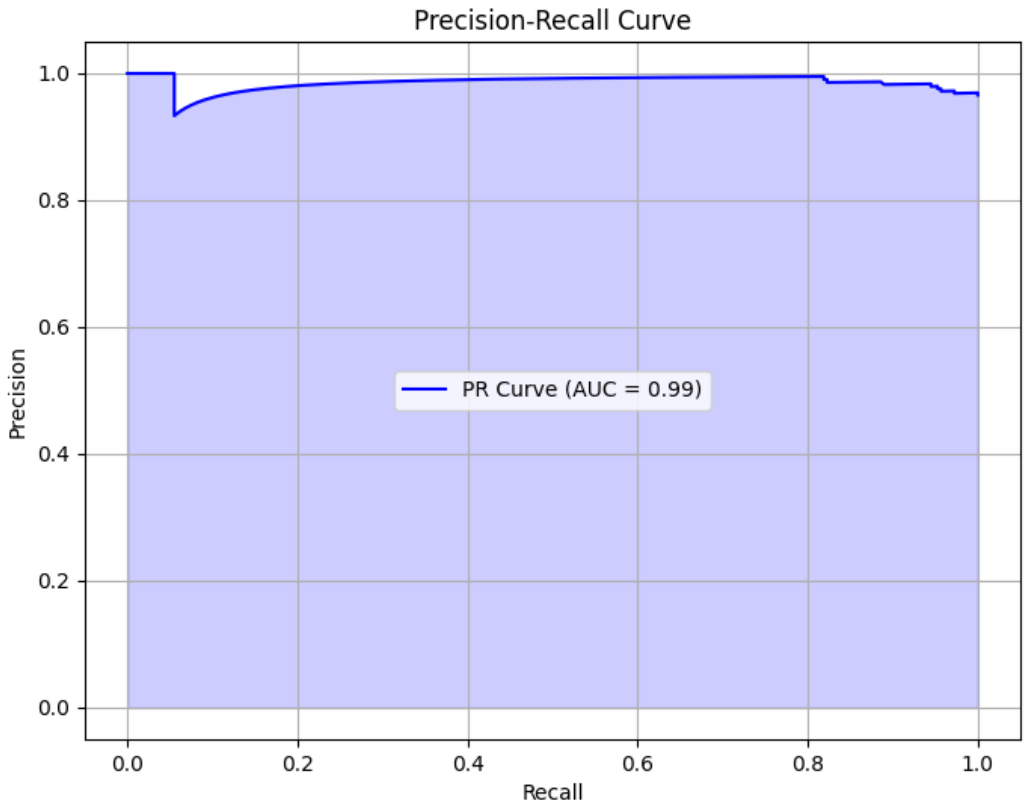


Figure 4. 20 PR Curve of Balanced Bagging Random Forest Model – Trained and Tested on Provisioned Dataset

```
ACTION
PROMOTED        1268
NOT PROMOTED      44
Name: count, dtype: int64
F2 Score: 0.97
PR AUC: 0.98
Confusion Matrix:
 [[  4   5]
 [  9 245]]
Classification Report:
                precision    recall  f1-score   support

NOT PROMOTED       0.31      0.44      0.36         9
    PROMOTED       0.98      0.96      0.97       254

    accuracy                           0.95       263
   macro avg       0.64      0.70      0.67       263
weighted avg       0.96      0.95      0.95       263
```

Figure 4. 21 Metrics of Balanced Bagging Random Forest Model with SMOTE-ENN – Trained and Tested on Provisioned Dataset

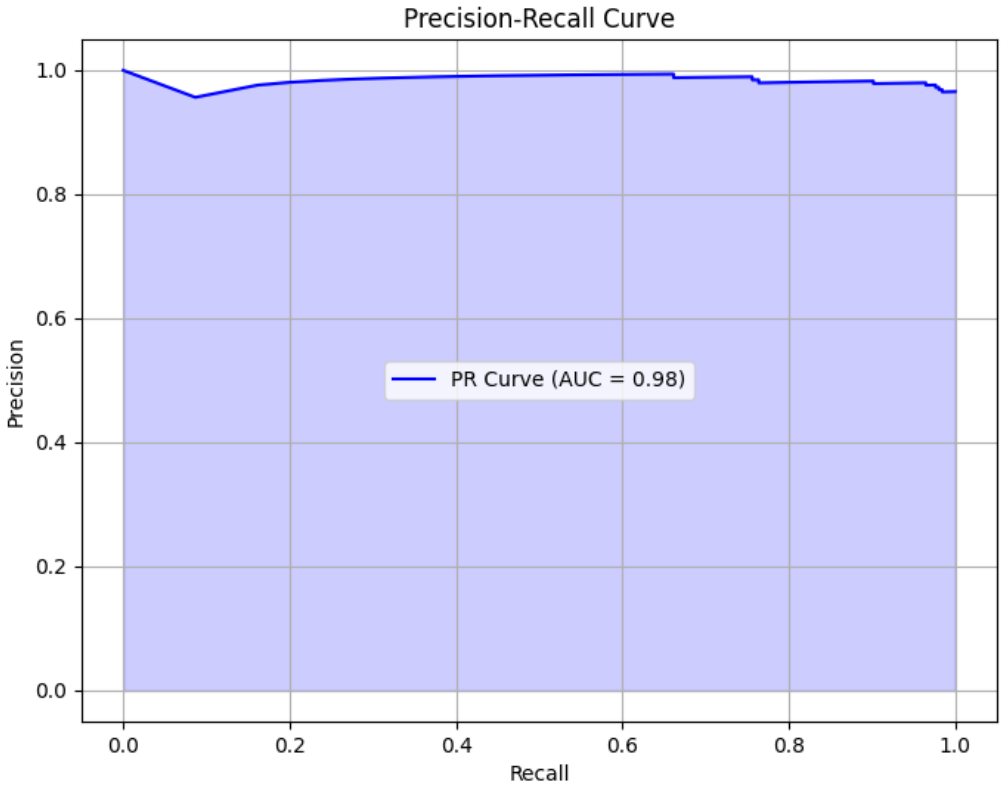


Figure 4. 22 PR Curve of Balanced Bagging Random Forest Model with SMOTE-ENN – Trained and Tested on Provisioned Dataset

4.3.3 Overall Evaluation

The following were the metrics recorded during the evaluation of the various models experimentally configured. The statistical results of the experimentation is shown in the table below:

Table 4. 3 Overall AI Model Evaluation

Metric	Baseline Logistic Regression (Control)	Logistic Regression	Random Forest (Base)	RF + SMOTEENN	RF+ Balanced Bagging	LogReg + SMOTE	RF + Balanced Bagging + SMOTEENN	Interpretation
F2-Score	—	0.81	0.98	0.86	0.78	0.84	0.97	Prioritises recall over precision; useful for catching more positives
PR AUC	—	0.99	0.98	0.98	0.99	0.99	0.98	Balance between precision & recall; more reliable than ROC AUC for imbalanced data
Accuracy F1	0.93	0.78	0.95	0.83	0.75	0.81	0.95	% of overall correct predictions; may mislead if classes are imbalanced
NOT PROMOTED Precision	0.94	0.11	0.17	0.09	0.11	0.13	0.31	Out of all NOT PROMOTED predictions, how many were correct
NOT PROMOTED Recall	0.99	0.78	0.11	0.44	0.89	0.78	0.44	Out of all actual NOT PROMOTED cases, how many were caught
NOT PROMOTED F1-Score	0.96	0.19	0.13	0.15	0.19	0.22	0.36	Balance of precision and recall for NOT PROMOTED class
PROMOTED Precision	0.78	0.99	0.97	0.98	0.99	0.99	0.98	Out of all PROMOTED predictions, how many were correct
PROMOTED Recall	0.25	0.78	0.98	0.84	0.74	0.81	0.96	Out of all actual PROMOTED cases, how many were caught
PROMOTED F1-Score	0.38	0.87	0.97	0.90	0.85	0.89	0.97	Balance of precision and recall for PROMOTED class
Macro Avg F1	0.67	0.53	0.55	0.53	0.52	0.56	0.67	Equal-weighted average of F1-Scores; good for seeing class balance
Weighted Avg F1	0.92	0.85	0.95	0.88	0.83	0.87	0.95	F1 average weighted by class size; good for overall performance

Key: Highest Metric Score
Final Selected Model Highest Metric Score

Note: Evaluation Metric Formulae Provided within Chapter 5, Section 2, Subsection 3.

The final model selected for implementation is the **Random Forest Model**, for its overall positive performance across various metrics in the model experimentation, and with

especial emphasis on the performance of the model with regards to the achieved minority class metrics.

4.3.4 Challenges Encountered

- AI Model Dataset Inadequacies
    - The training of the AI Model was not performed with optimal conditions. The sample size of 1312 records may have been sufficient for the training of the classification model; however, it was revealed that the dataset possessed great imbalance in the target classification. These imbalances had to be mitigated through the selection of an appropriate model algorithm and other supplementary functions, such as SMOTE (Synthetic Minority Over-sampling Technique) - a method that creates synthetic data to address class imbalances.

4.4 Main Functions, Models, Frameworks

4.4.1 Introduction

This section outlines how the developed system fulfils the research objectives through a series of integrated modules, models, and frameworks, each carefully designed to address core challenges and goals associated with automating the processing of HR case documents.

4.4.2 Alignment with Objectives

4.4.2.1 Objective 1

“To identify and analyze the primary challenges in developing an AI-driven system for HR case processing.”

The primary challenges addressed in this framework were:

a) Impacting factors to OCR accuracy: Tackled by implementing a wide-coverage Image Pre-processing Module that includes grayscale conversion, noise removal, skew correction, font thickness adjustment, and border removal.
b) Wide range of diversity in HR case types and document layouts: Managed by developing a Region of Interest (ROI) Detection Module to dynamically identify and isolate relevant areas of text on varied document layouts.
c) Domain specific language and sentiments: A custom lexicon-based Sentiment Analysis Module was created to interpret extracted text within the HR domain, focusing on subjective elements such as cues indicating potential acceptance or rejection skews.

d) Greatly imbalanced datasets for machine learning classification: The AI Classification Module employed techniques such as Synthetic Minority Over-sampling Technique (SMOTE), Edited Nearest Neighbour (ENN) and class balancing to handle imbalance in the dataset used for promotion prediction, in order to minimize the effect of the imbalance on the developed model classification functionality.
e) Scalability and adaptability for real-world deployment: The entire framework was developed with modular architecture using Python, enabling scalability, iterative debugging, and independent module enhancement. Scalability is as well achievable due to the relatively low functional requirements and selection of technologies that are easily scalable, such as MySQL for scalable data handling.

These challenges informed both the design logic, and the selection of technologies and algorithms integrated into the system.

4.4.2.2 Objective 2

"To design and develop a system that integrates AI, OCR and custom lexicon-based sentiment analysis (SA) to automate the processing of HR case documents."

This objective is fulfilled by the development of an end-to-end modular framework that cohesively integrates:

a) Computer Vision (OpenCV) for all image-related operations, from importation to text region detection.
b) OCR (Pytesseract) for extracting textual content from scanned HR documents.
c) Custom Lexicon-based Sentiment Analysis embedded within the OCR pipeline, using regular expressions and HR-specific lexicons to identify tone and make soft recommendations.
d) AI Classification (RandomForest Balanced Bagging with SMOTE-ENN Classifier) for predictive tasks, such as determining promotion likelihood based on case features.

Each module feeds data into the next via clearly defined data pipelines, enabling seamless transformation from raw scanned images to structured output data stored in a MySQL database, demonstrating true integration of multiple AI-related techniques.

4.4.2.3 Objective 3

"To facilitate the accurate extraction of key information from scanned HR case documents using RegEx."

This was accomplished by incorporating custom-developed regular expressions (RegEx) within the OCR Module, allowing:

a) Structured parsing of freeform text directly extracted from documents or using ROIs.
b) Precise identification of recurring HR case data points such as employee Identification, decision dates, eligibility indicators, pay scales, and case assessor.
c) Preprocessing of extracted text into usable structured format, feeding into subsequent sentiment, prediction and data storage modules.

This targeted approach outperforms generic extraction by being highly tuned to the specific semantics of the HR case domain.

4.4.2.4 Objective 4

"To evaluate the performance of the developed system in terms of its accuracy, efficiency, and scalability, in comparison with traditional manual HR case processing methods."

While detailed evaluation and comparison with manual methods is provided in **Chapter Five**, the framework design does incorporate performance and reliability features:

a) Accuracy: Achieved through an enhanced OCR pipeline with visual feedback and validation, alongside high-performing sentiment analysis logic. The AI classifier evaluation produced an F2 score of 0.97, a PR AUC score of 0.98 and a macro average F1 score of 0.67- the highest score achieved of all tested models and equivalent to the achieved score of the control model. Additionally, the OCR accuracy rate achieved is greater than 95%.
b) Efficiency: Achieved by automating formerly manual stages of case processing - including text identification, extraction, and sentiment classification - significantly reducing human workload. Case processing speed improved significantly- from 3 days in manual contexts to less than 10 seconds for a two-page case batch, inclusive of case digitization, or 3.3 seconds for a batch of 1000 records using the integrated AI model solution
c) Scalability: Promoted through modular Python architecture and database integration, allowing deployment in more complex or higher-volume environments with minimal reconfiguration as well as the incorporation of an easily scalable database solution.

In combination, these features validate the framework's viability as a scalable and effective AI solution for HR departments.

4.4.2.5 Objective 5

"To provide recommendations for the future refinement of integrated machine learning models, analysing methods with which the limitations of the technologies and their general augmentation may be achieved."

Detailed recommendations for future work and improvements to the proposed solution are indicated in Chapter 6 of this research, however some recommendations are highlighted in brief below:

a) Utilisation of Technologies that mitigate the adverse effects of data imbalances, to achieve the most optimal model results. ML Model training using free-form unstructured text: Utilising other more advanced and functionally diverse models such as BERT, fine-tuned for the processing of HR data for deeper understanding of more nuanced, complex cases.
b) Expansion upon the size of the training dataset either through the digitisation of images for their conversion into structured records or through the provision of additional data or datasets by other governmental institutions with similar structured procedures and data formatting.
c) Deeper integration of the ABSA and AI Module functionalities: The intentional combination of the SA and ML Model technology may yield greater benefits and opportunities for modular scaling than the exclusive use of either of the solutions to accomplish decision support through predictive recommendation. This approach may also promote greater explainability of recommendations.
d) Add structured features: Improved training on larger, more balanced datasets with more features (columns) to allow greater refinement towards the predictive functionality of the integrated AI Model.

4.5 Chapter Summary

This chapter outlines the comprehensive modelling and development framework employed within the project, focusing on the digitisation of HR case documents. The modelling approach integrates technologies various technologies, including OCR, regular expressions, machine learning based AI classification, and sentiment analysis (SA) to transform raw scanned documents into structured data stored in a MySQL database, while providing decision support to users through either SA or the integrated AI-based alternative. Advanced image processing techniques, such as those implemented utilising OpenCV, optimise document scan quality for OCR, addressing challenges like skew, noise, and misalignment, thus improving text extraction

accuracy. The experimentation with and addition of an AI Model for the classification of cases, yields great potential as an approach to automation and efficiency increase within the domain. The integration of Python tools ensures modularity and scalability, making the solution adaptable for varied HR document processing needs.

The project utilises the agile development methodology, enabling iterative improvements and feature enhancements. Key functions implemented include image pre-processing, extracting of text and structured data using OCR and RegEx, and the conduction of aspect-based sentiment analysis for decision support. In parallel, an AI classification model has been developed to serve as an evolutionary enhancement or alternative to SA, offering a data-driven predictive capability for outcome suggestions. The structured data is then consolidated and stored in a scalable MySQL database, ensuring seamless integration and accessibility. MySQL's robustness and OpenCV's advanced image processing capabilities are highlighted as essential to the system's efficiency and functional reliability. A custom lexicon for ABSA allows domain-specific sentiment evaluation, tailored to HR document processing requirements, particularly in governmental contexts, where document structure is held consistent.

The framework demonstrates a thorough, cohesive pipeline that begins with raw image scan input, processes the data through several modular steps, concluding with the storage of actionable case data for retrieval and reference. Designed for efficiency improvements to traditional manual document processing methods, as well as scalability, and future integration of advanced technologies like customised or otherwise suitable machine learning models, the project provides a powerful and adaptable solution for automating HR case document processing.

As part of this forward-looking integration, a functional AI model based on machine learning principles has been implemented to complement and potentially replace the sentiment analysis module. Adaptable to using the structured data outputs from the OCR-RegEx pipeline, the model applies advanced classification techniques (e.g., Logistic Regression, Random Forests and XGBoost), trained on labelled historical HR decision data to predict promotional HR case outcomes. The model incorporates robust pre-processing pipelines, stratified sampling, class imbalance handling, and performance evaluation using metrics like Precision-Recall AUC and F2-Score. This AI-driven component introduces a data-centric alternative that may offer improved consistency and predictive accuracy compared to rule-based sentiment methods.

This AI solution signifies an evolutionary step beyond traditional SA by introducing a contextually trained classifier that learns from nuanced patterns in decision data, including factors like age, location information, and position. By integrating this AI model into the existing modular architecture, the system becomes capable of dual-mode decision support—rule-based via sentiment scores and pattern-based via predictive analytics—enhancing its utility, interpretability, and long-term adaptability for HR case analysis in public sector environments.

The next chapter will explore the detailed findings of the research, exploring solution evaluations and analysing the results of the research's development.

CHAPTER 5 – RESULTS AND DISCUSSIONS

5.1 Results Presentation

5.1.1 Introduction

The development of the automated HR case document processing system showcases a significant step in the utilisation of modern, revolutionary technologies for a novel application with the purpose of enabling document digitisation, and decision support. Leveraging AI and OCR technology to incorporate automation-based efficiency and convenience into moder day HR department workflows in governmental contexts. The vital aspect of document digitisation, allows the system to be built tailored to the parameters of a specific document type, potentially providing a means to eliminate digitisation backlogs when applied on a greater scale. This project presents the development of a framework of technologies for the improvement of HR case document processing. Developed within a timeframe of 3 months, the solution prioritizes scalability, replicability, and accuracy while maintaining low computational overhead.

5.1.2 Overview of Results

The developed solution was completed within a timeframe of 3 months. Results presented by the system through testing indicate effectiveness for the application for which it was designed. The system enables scalable and replicable document processing and swift document digitisation. The solution was developed within Python and involved the integration of numerous technologies in a coalescent manner to achieve the overall goal of designing a Framework for automated HR case document processing using Artificial Intelligence and OCR.

Development, testing and result data gathering were conducted on a single laptop computing device with the following performance and environmental specifications:

- Processor: Intel i7-7700K Processor @ 4.2 GHz – 4.5 GHz.
- Memory: 16.00 GB DDR4.
- Operating System: Windows 10 Pro, 64 Bit
- 1 TB HDD / 320 GB SSD (note: Sufficient disk drive space was maintained throughout the workings of the project so as not to impact performance testing)

The development and testing environment enabled the completion of the development and testing tasks, and establish a measurable performance to specification expectation. The project solution, is recommended to ideally be implemented on computing devices with higher

Processor specifications, and with the use of a Solid-State Drive if available. However, the solution, remains a relatively lightweight application with low minimal running requirements.

While the image pre-processing functionality possesses relatively low computational performance and time overheads, with average runtimes of image transformation operations being completed within 1-2 seconds of initiation of the process for all image pre-processing steps (with an addition of 4 seconds whenever a recompilation of the project is performed), the OCR, ABSA and RegEx pattern recognition functionalities, require significantly more processing time, measured at an average of 5.40 seconds for all actions including insertion of processed records into the associated MySQL database. The AI module analysis processes case batches with similar efficiency with a batch of 1000 records requiring an average processing time of 3.3 Seconds. Overall, despite the processing time required, the benefit offered is still considered worthwhile. It is as well theorised that computational environments of higher specifications are likely to greatly shrink processing time requirements.

It is imperative for the sake of understanding that a control variable be present when considering timings of workflow execution, in order to ascertain how much value truly is added by the developed framework. HR professionals directly involved with the scope and duties of the work currently being automated have been consulted to gain feedback into the processing time of a case from the stage at which the document that is analysed is produced. Once the document is produced, it is presented to 3 subsequent levels of scrutiny and approval, followed by a consensus-based approval process involving the rotation of the document across multiple offices. This process is a time consuming one that may take on average 3 days, and potentially result in reversion- further extending the processing timeline, before further escalation, until reaching the state of consensus-based approval. The time savings of the system are thus highly relevant. Enabling the digitisation of documents for simultaneous access across relevant levels at the appropriate time, and supporting the provision of feedback at earlier points in the workflow.

OCR produced a measured accuracy rate of 95% where the scanned text consists of non-special alphabetic characters and numbers. Font selection may greatly improve or worsen OCR performance. Fonts with distinctive lettering and numbering are more likely to produce higher accuracy results. OCR suffers significant accuracy loss when text is wrapped around images or images overlap text. Single column documents yield higher OCR accuracy rates than multi-column documents. Lettered listings with left alignment are accurately identified but retrieved in incorrect locations, it is hence recommended that documents meant for OCR

processing are single column format with minimal to no left or right aligned list denotation characters such as bullet points, roman numerals, bracketed letters, etc. Should the recommendations be adhered to, a 95% OCR scan accuracy rate is feasibly achieved. Overall, OCR scan performance has been satisfactory for the purposes of this project, despite minor positional errors, and character misreads.

Regex match retrieval had achieved an accuracy rate of 80% across all matches. This rate reduction is due to scenarios involving the processing of mixed case types and case types with similar text presentation conditions which lead matches to be placed in the wrong match containers. Additionally, the regular expression finditer and findall methods do not allow for regex matches that contain lookaheads, due to their frequent causing of zero-width insertion related errors. This meant that certain matches could not be executed and workarounds that were not optimal had to be employed to receive meaningful data.

The cv2 display function is used to create windows and define their sizing dimensions as well as interactions, in order to view the changes that the images undergo as they are prepared for further processing through the integrated Tesseract OCR engine. Each process in the pipeline was visually checked in order to determine error points and inaccuracies for refinement of the system.

In line with the achievement of the project aim and objectives, the developed solution seems to be in alignment with the overall goal of enhancing efficiency and accuracy of HR case processing by leveraging OCR, custom lexicon-based ABSA and AI to automate the analysis and processing of HR case documents.

In relation to manual, paper-based work practices, the system achieves its goal of reducing on the inefficiencies, errors, and delays that watermark manual, paper-based systems and allowing for improved decision making with regards to Human Resource Cases.

The developed solution satisfies Objective 1 and 3, and allows for the further satisfaction of Objectives 2 and 4 through performance benchmarking and comparison with manual alternatives. Objective 5 has been satisfied through the workings of this research report which speaks to alternative AI based models for sentiment analysis, such as BERT and VADER, with additional insights drawn from extensive research materials including the Semantic Evaluation Workshop Series conference papers within which ABSA technologies and advancements thereof are greatly focused on and refined through problem solving challenges. Thus, this project and its produced system, satisfy the set objectives and aim.

5.1.3 Summary of Primary Outcomes of the Project

Development and Deployment of Solution:

- Successfully achieved the development and testing of the proposed automated HR case document processing system within an estimated timeframe of 3 months.
- The system leverages the technologies of OCR and ABSA for the enhancement of performance accuracy and efficiency.

Performance and Efficiency:

- Achieved a scalable, replicable document digitisation solution with minimal computational overhead.
- All sequenced image pre-processing operations measured for completion in average range of 1 to 2 seconds, with a 4 second addition when a recompile is performed prior to execution. Additionally, full document processing, including database insertion, averaged an execution time of 5.40 seconds, with AI model-based processing of a 1000 case record batch requiring 3.3 seconds.
- The lightweight system requirements facilitate adaptability, though higher specification computing systems are recommended for improved performance.
- Presents an OCR accuracy rate of 95% for non-special, alphabetic characters in single column format.
- Regex match retrieval achieved an accuracy rate of >85% across all matches.

Alignment with Objectives:

- Fulfilled all stated project objectives, particularly with regard to the reduction of inefficiencies, error occurrences, and delays in manual HR case processing.
- Achieved demonstrable improvement in decision-making capabilities through document analysis and processing workflow automation, for improved decision support.

System Capabilities:

- The solution integrates a coalescent framework of Python and other supporting technologies to deliver a refined, reliable and, modular solution.

- System is recommended for deployment on higher-specification computing systems than that of the development and testing environment. However, testing reveals that performance requirements of the developed system are relatively low.

Impact and Practical Benefits:

- Significant reduction in processing time and operational inefficiencies compared to manual, paper-based systems.
- Elimination of need for manual database entry, as data is digitizable at earlier stage of workflow, and utilises documents already existent in the workflow.
- Poised to present as a scalable and effective augmentative tool for HR departments and personnel to enhance productivity and accuracy.

5.1.4 Main Achievements of Solution Approach

The developed automated HR case document processing system represents a significant step forward in leveraging Artificial Intelligence (AI) and Optical Character Recognition (OCR) technologies to enhance efficiency and accuracy in HR workflows. The following outlines the key achievements of the solution approach:

1. Successful Development and Implementation

Completion Time: The solution was developed and tested within an estimated 3-month timeframe.

Technology Integration: The system, built within the python programming environment, integrates with AI, OCR and other technologies for image and text processing to create a cohesive framework, culminating into a system with a balance of technological synergy for the purpose of addressing the nuanced complexities of HR document processing.

2. System Performance and Scalability

Efficient Document Processing: The system presents a scalable, replicable solution for document digitisation with minimal computational overhead.

Image Pre-processing: Operations average time is 1-2 seconds per image collection (tested collection sizes being 2 pages on average), with a slight increase of 4 seconds for scenarios involving recompilation.

Full Workflow: Document processing, inclusive of OCR scanning, text pre-processing, regex match extraction, variable preparation and database insertion, averages 5.40 seconds per complete instance.

Lightweight Requirements: The developed solution runs with relatively low performance requirements, making it feasible for application when higher performance computational devices are unavailable, though it is effective on lower performance computing devices, higher-performance systems are recommended for optimal results.

3. High OCR Accuracy: Achieved a 95% OCR accuracy rate for non-special alphabetic characters in single-column documents. HR case documents considered within the development of the system have been primarily single column documents, this allowing for a high consistency in accuracy rates. Additional guidelines can however, be established to enhance performance further through establishment of further standardised document formatting, according to observations of project testing.

4. Fair Regex match extraction accuracy rate: The Regex match extraction accuracy was measured at a rate of 80% with errors primarily caused by overlapping match rules resulting in a few matching placement issues.

5. Alignment with Project Objectives: The system fulfils all stated objectives by: Reducing upon current process inefficiencies, errors, and case processing delays associated with manual HR case processing methods.

6. Practical Benefits and Impact on Domain

Improved Efficiency: The system significantly reduces processing time compared to manual process workflows, provides decision support and eliminates the need for manual database entry.

Ease of adoption in varied environments: The lightweight design offers a major advantage, allowing the system to be applied in situations and scenarios where available computational power, may be low. The entire workflow being compacted into one running system ensures ready availability of the efficient system.

Enhanced Productivity: The automation of repetitive tasks presents an opportunity to allow HR personnel to focus on more pivotal and nuanced tasks requiring human intelligence and reasoning, as well as their professional domain specific knowledge.

7. Framework Design

Modularity and Reliability: Python is able to use a diverse set of modules and tools that enable the practical development and implementation of AI oriented applications. The framework integrates seamlessly with the additional tools, allowing for extension of functionality as needed in further iterations of the produces system, thus allowing greater flexibility and agility in development and refinement of the solution.

Error Monitoring: Real-time process visualization using cv2 for the display of images as well as other useful debugging tools allows for prompt identification and resolution of errors during development and testing.

5.2 Analysis of Results

5.2.1 Introduction

The previous chapter elaborated on the results of implementation and testing of the framework, highlighting positives as well as drawbacks and areas for improvement. This section will thoroughly analyse the results retrieved from the testing and implementation of the project in order to determine benefit of application and integration for its designated purpose and ascertain whether further development and refinement is warranted.

5.2.2 Recap of Project Objectives and Status

A detailed recap of the project objectives and their status in terms of fulfilment or partial fulfilment is found in chapter 4, section 4, subsection 2. In brief it is worthy of note that all set objectives have been fulfilled.

5.2.3 Interpretation of Key Results and Identification of Strengths

5.2.3.1 Performance Metrics

Metric 1: Accuracy Percentage of OCR Results

Achieved result: 95%

Context: While text OCR accuracy may greatly vary when handling handwritten text, printed text has much higher average accuracy rates. The Tesseract engine and others average above 95% accuracy in the scanning of printed text [81][82].

Metric 2: Processing Time per Cycle

Achieved result:

- For Image Scans: 5.40 seconds (Text Processing, ABSA, Data Insertions) + 4 Seconds (Image pre-processing) = 00:09:40 minutes total execution time.

- For Batched Records(Sample Size of 1000 Records) : 3.3 seconds (Selection of document, Processing of Case Batch, Production of Processed Case Batch Document with Decision Recommendation s and certainty score)

Context: Manual Processing may take on average 3 days to process a document and arrive at a decision in normal workflow routines.

Metric 3: Resource Utilisation

Table 5. 1 Resource Utilisation Metrics

Task	CPU Usage	RAM Usage	DISK Usage
Python (Idle)	0.1064GHz	6,080KB	152B/s
Image Pre-Processing	0.79 GHz	270,438.4KB	12,500B/s
Tesseract (OCR)	0.63 GHz	120,652 KB	n/a
Discussion: As observable the performance requirements for the operation of the developed solution are relatively low. Recommended specifications for optimal performance are: 2-4GB RAM, 1.5GHz CPU speed, Windows 10 or later.			

Metric 4: AI Model Metrics

Table 5. 2 Confusion Matrix Interpretation

	Predicted: Positive	**Predicted: Negative**
Actual: Positive	**TP** (True Positives)	**FN** (False Negatives)
Actual: Negative	**FP** (False Positives)	**TN** (True Negatives)

Table 5. 3 Metric Formulae

Metric	Formulae
F1-Score	$$\text{F1 Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$
F2-Score	$$\text{F2 Score} = 5 \cdot \frac{\text{Precision} \cdot \text{Recall}}{4 \cdot \text{Precision} + \text{Recall}}$$

Precision	$$\text{Precision} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Positives (FI}}$$
Recall/Sensitivity	$$\text{Recall} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Negatives (FN)}}$$
Weighted Average F1	$$\text{Weighted F1} = \sum_{i=1}^{K} \left(\frac{n_i}{N} \cdot \text{F1}_i \right)$$ • n_i: number of true instances for class i • N: total number of instances • F1_i: F1 score for class i
Macro Average F1	$$\text{Macro F1} = \frac{1}{K} \sum_{i=1}^{K} \text{F1}_i$$ • K: number of classes
PR AUC Score	$$\text{PR AUC} = \int_0^1 \text{Precision}(\text{Recall})\, d\text{Recall}$$

Table 5. 4 AI Model Metrics

Metric	Score	Interpretation
F2-Score	0.97	Prioritises recall over precision; useful for catching more positives
PR AUC	0.98	Balance between precision & recall; more reliable than ROC AUC for imbalanced data
Accuracy F1	0.95	% of overall correct predictions; may mislead if classes are imbalanced
Minority Class Precision	0.31	Out of all NOT PROMOTED predictions, how many were correct
Minority Class Recall	0.44	Out of all actual NOT PROMOTED cases, how many were caught
Minority Class F1-Score	0.36	Balance of precision and recall for NOT PROMOTED class
Majority Class Precision	0.98	Out of all PROMOTED predictions, how many were correct
Majority Class Recall	0.96	Out of all actual PROMOTED cases, how many were caught
Majority Class F1-Score	0.97	Balance of precision and recall for PROMOTED class
Macro Avg F1	0.67	Equal-weighted average of F1-Scores; good for seeing class balance
Weighted Avg F1	0.95	F1 average weighted by class size; good for overall performance

Contributing Factors

Factor 1: Standardised document format and formatting guidelines

Contribution: The standardisation of document formatting allows regex matches to consistency find the data that they are programmed to find, promoting consistency and replicability. The addition of guidelines allows the elimination of document formatting practices which may affect the quality of OCR results of regex matches.

Factor 2: Workflow Design

Contribution: The workflow design allows the execution of the programme from start to finish within the data pipeline in 00:09:40 min, and allows for pre-processing, data storage and decision support within each execution

Key Strengths

Strength 1: High OCR Accuracy Rate (95%)

Evidence: Assessment of the error margin with regards to OCR. Largely, non-alpha numeric characters were the source of accuracy reduction. However, these characters did not affect accuracy percentage a great deal with clean single column paragraphs.

Strength 2: Low Performance Requirements and Customisability

Evidence: Test results indicate the performance requirements of the system being relatively low, this system can thus be easily introduced in many environments, and additionally is able to be tailored to specific domain needs, as long as the consistency of document structure is maintained.

Strength 3: Modularity

Evidence: The development environment (python), allowed for the easy integration of modules, usable for various functions. Functions which may be optional or mandatory in the workflow as required, and allow for new functions to be introduced at various stages of the pipeline without significant performance overheads.

5.2.4 Summary of Key Insights

Performance Metrics:

OCR Accuracy: 95% accuracy rate for printed text, typical of top OCR engines.

Processing Time: 00:09:40 minutes, more optimal than manual processing (3 days).

Resource Utilisation: Low CPU, RAM, and disk usage, suitable for low-end PCs.

Performance Requirements:

Minimum: 2-4GB RAM, 1.5GHz CPU, Windows 10 or later.

Low system requirements make the solution adaptable to various environments.

Standardised Document Format: Enhances OCR and regex consistency.

Workflow Design: Efficient processing from start to finish.

5.3 Comparison to Related Work

5.3.1 Introduction

The previous section contains an analytical discussion on the performance results and findings obtained from the development of the project and its testing. This section provides a

comparative analysis of the technical aspects of and emphasised by existing works in relation to this project's aim of implementing an automated HR case processing using OCR and ABSA technology. The discussion highlights technological alignments and divergences, that situate this project within the broader field of AI HRM applications. Through this reflective exploration of other projects in the field, introspective perceptions may be received on the value of the projects and factoring aspects thereof.

5.3.2 Comparison to related works

5.3.2.1 AI for Increased Efficiency in HRM

Reference [1] and [24] discuss the role of AI in improving efficiency by automating repetitive HR tasks such as recruitment and performance evaluations. By creating a solution that provides efficiency improvements within the context of HR document processing and as well with that solution achieving fair timing benchmarks (1:02:30 min total processing time) and low operational requirements; efficiency improvements have been reasonably realised. Additionally, research revealed that AI provides capable solutions for HR professionals, by handling time consuming repetitive tasks, and eliminating bias in said tasks [83].

5.3.2.2 Document Digitization and OCR Integration

In alignment with the highlight of reference [9] on OCR's utility in the digitising of documents, this project has successfully implemented automated document digitisation through the use of OCR in conjunction with RegEx for extraction of desired, key data points. Reference [4] explored the integration of OCR with Robotic Process Automation (RPA). The concept of RPA aligns strongly with the workflow digitisation process, provisioning an alternative process to manual completion of tasks by humans, this with special regard to automating repetitive tasks that occupy valuable HR professional time.

This project leverages the researched insights and succeeded in achieving a 95% OCR accuracy for printed text OCR extraction. The developed solution addresses compatibility and adaptational concerns by functioning with relatively low computational requirements, such as a recommended random access memory requirement of 2–4GB, with a processor speed of 1.50GHz.

5.3.2.3 Sentiment Analysis for Decision Support Systems

According to [39] and [40], Sentiment analysis has been widely applied within organisational decision making. This project leverages Sentiment Analysis in an effort to provide decision making support to HR professionals through identified and sentimentally

scored aspects, weighing towards an aggregated score and overall sentiment. This project employs lexicon-based ABSA to offer actionable insights.

5.3.2.4 Human-AI Collaboration in HRM

The earlier analysed research references [18] and [45] embraced the notion that AI must be used to complement human decision-making, and not replace it. This intent is well reflected in the project, with the solution designated to augment human abilities, requiring and embracing human engagement in work processes. AI professionals possess pertinent domain specific knowledge, which allows the making of decisions in highly nuanced matters, where AI applications in their current state may not be able to provide appropriate decisions.

5.3.2.5 Advanced Applications of AI in HR

Reference [19] and [36] discuss the potential for intelligent systems to automate complex task execution. These studies provide a theoretical basis for the integration of AI within this project with the purpose of enhancing decision timeliness and operational efficiency. By achieving low-resource adaptability and integrating scalable technologies, the project advances the application of AI in the HRM domain, within the specific context of government HR case processing.

5.4 Unique Contributions of this Project

This project explores the automation of Human Resource Case Processing within primarily paper-based Governmental contexts with the added beneficial process of document digitisation with a focus on the elimination of paper-based record backlogs. The research identified gaps in existing applications for Human Resource domain contexts, leveraging on the identified opportunity to apply existing technologies in a novel manner to accomplish the digitisation of documents and automation of case processing through provision of decision support; Creating a framework integrating, with coalescence: Image Processing, Natural Language Processing, and Artificial Intelligence to successfully solve the inefficiencies of the defined domain.

This project uniquely contributes a novel, modular framework. for automated processing of scanned HR case documents within a governmental or bureaucratic context. Unlike many existing solutions that treat OCR, sentiment analysis, and AI classification as discrete efforts, this project brings together these components into a cohesive pipeline tailored to structured HR documentation. The design leverages the consistent structural format of government documents to optimize OCR performance, achieving a reliable accuracy rate of 95%. The project further introduces a sentiment analysis submodule, using a custom-built

lexicon specific to HR terminology, which serves as both a functional intermediary for case assessment and was implemented as a precursory approach to the subsequently implemented AI-based decision support.

Significantly, the AI classification module is the result of iterative experimentation with multiple model types and balancing strategies to address the challenge of severe class imbalance in real-world institutional data. The implementation and evaluation of the Random Forest Classification Model with Balanced Bagging and SMOTEENN stand out as a critical contribution, offering a robust, context-aware predictive model. This model not only performed strongly on standard metrics such as the F2-Score and PR AUC but also maintained meaningful sensitivity to the minority class—a common shortcoming in imbalanced classification tasks.

The integration of this model into the processing pipeline represents a practical, deployable solution for digitisation and decision support in environments where manual HR processing remains the norm. Additionally, the system's performance—completing a full case batch process in under 10 seconds—demonstrates its capability to deliver operational efficiency without dependence on high-end computational resources. This project contributes to the body of knowledge by bridging OCR, sentiment analysis, and AI classification within a practical, domain-specific implementation, supporting the development of intelligent automation in sectors traditionally underserved by such technologies.

5.5 Implication of Results

5.5.1 Introduction

The previous section discussed the technical and technological alignments of this project with related and similar works. This section speaks to the greater implications of the results of the research. This research presents an opportunity for the refinement of governmental HRM practices in a novel manner, by implementing a theorised and practically developed framework for human resource case processing. This project seeks to make a lasting and foundational impact, paving the way for more research exploration into the domain.

5.5.2 Technological Implications

Several technologies were integrated into the framework, each necessary for the functional options and opportunities provided. One fundamental technology is the utilisation of image processing module: cv2. The cv2 module contains various functions for advanced image processing, and allows for the manipulation of images as may be necessary for preparation for processing. Matplotlib, another key module which was utilised in conjunction with the cv2 module, allows for the plotting of bounding boxes through image processing and

coordinate mapping over blurred segments to identify regions of interest for OCR, which may then optionally be used to conduct targeted OCR operations.

To extract text from images, an OCR engine was necessary. Tesseract OCR is an open-source OCR engine which has a python module for seamless integration into projects. The tesseract model served as the means with which text was extracted within the data pipeline. Images once pre-processed are subjected to the OCR extraction function, which uses iteration to ensure that all pages of a document collection input into the system are scanned and text extracted. The text can then be segmented to ease its processing. RegEx, is another technology that was integrated into the project for its proficiency to manipulate text in a fine-tuned manner. Patterns may be sought out to determine boundaries and starting points for meaningful text, others may be set for the extraction of key variables such as file codes and personal details within cases.

Once useful input is extracted through regex patterns and iteratively stored into list or dictionary variables as may be optimal. Simultaneously another technology is implemented into the framework, for the creation of a lexicon (dictionary), containing sentiment and score designations for phrases. This function scans paragraphs prepared by the OCR and RegEx functions and identifies sentiments expressed within each segment scanned. A segment is representative of a single case.

Once the sentiment is retrieved for each segment, and stored into a dictionary, all data is now ready for insertion into a linked database. MySQL was chosen as the database solution as it presents scalability opportunities and coalesces with multiple technologies with ease and efficiency. The mysql.connector module is imported to enable the use of functionality to enable establishment of a link between the system and the created database for system data. Insertions are then performed of all the relevant data and the data pipeline concludes.

This technology combined in this novel way provides a robust framework for Human Resource case processing, which is scalable and adaptable to changing requirements and scenarios. This projects application of these technologies presents a novel, precursory approach to further refinements and applications of the technology in future works.

Project testing revealed that the program satisfied reasonable benchmarks for processing speed, and performance requirements as well, making this a diversely feasible solution with lightweight implementation.

5.5.3 Practical Implications for HR Applications

The results of the system testing and implementation indicate the potential for significant benefit when implemented, and integrated into HR department workflows. Automated human resource case processing enables the diversion of Human Resource personnel to varied, more nuanced tasks requiring human expertise, whilst routine, repetitive tasks, such as data entry may be taken up by a system such as this and executed with reduced error. Additionally, with the use of ABSA and AI for the provision of case determination recommendations, based on presented case facts, time savings stand to be made. This provision of augmentative decision support, may greatly boost efficiency within HR departmental workflows.

HR departments may leverage the solution with easy implementation. Given the low computational requirements of the system, integration of the system shall be completed in minimal time. The system shall also present the opportunity to digitise documents and facilitate the dismantling of document digitisation backlogs, in tailored manner and in accordance with the needs of the department and workforce.

5.5.4 Ethical and Organisational Implications

To address ethical concerns revolving around the results and implementation of this project, certain factors must be considered in a deliberate manner. AI is a technology that may be subject to bias when model training is involved, models inherit biases found in data unless steps are taken to ensure that bias is eliminated. Within the context of this project, no ML models are integrated for the processing of data. However, to handle other forms of bias, rules and programming logic can be applied to ensure that bias is avoided or factors of bias have no bearing on the outcome of a case. By setting deliberate logical rules to prevent bias the system can be implemented in a fair, ethical manner.

HR professional displacement may be another area of ethical concern with the development and implementation of this project. However, this programme is intended for function with and alongside HR professionals, and is meant to act as an aid towards greater operational efficiency rather than a replacement for personnel.

In order to ensure that personal data is correctly processed and protected, applicable legislation within the locale of the solutions application must be consulted. This shall ensure compliance with stipulated data protection and privacy laws within the system.

5.5.5 Implications for Future Research and Development

Opportunities for further research and refinement abound within the project domain. Potential improvements to OCR scan accuracy or result processing logic could be implemented. More complex rules could be applied to solve more nuanced problems within the flow of work. Machine learning models may be integrated to provide case determination recommendations of a higher quality and with greater efficiency. More types of documents other than those within the scope of work could have customised logic integrated into them to enable dynamic processing of various case types based on various rules and regex pattern matches. The technology has several prospective angles of development and further research. This project seeks to set a foundation for the extension and improvement of the workflows involved.

5.6 Chapter Summary

This chapter presents the successful implementation and subsequent evaluation of an automated HR case document processing framework. It details the processes of the systems development, testing, and key performance metrics, including achievements in OCR accuracy rates, and regex match accuracy rates, with an average processing time of just over a minute per document. The lightweight and modular framework effectively reduces upon workflow inefficiencies, eliminates manual database entry and the need thereof, and supports decision-making through ABSA and AI. All project objectives were successfully met, demonstrating the scalability, adaptability, and practicality of the developed solution. The chapter concludes with a consideration of the implications of the results of this work.

CHAPTER 6 – SUMMARY AND CONCLUSION

6.1 Summary of Main Findings

The research successfully demonstrated the integration of OCR with ABSA and AI for the purpose of processing HR case documents, efficiently. Key result findings indicate the system's ability to automate text extraction in general and targeted manner, perform keyword identification, and sentiment evaluation with a high degree of accuracy. The application of regular expressions, additionally, ensured structured data extraction, enabling digitisation of document case record information and provision of decision support. The system has notably reduced on processing time and error rates in comparison to manual methods of data entry and task completion, showcasing its potential for governmental HR departments.

6.2 Contribution to the body of knowledge

This project contributes to the body of knowledge by addressing a gap in the application of mature technologies—image processing, OCR, natural language processing, and machine learning—in predominantly paper-based governmental HR contexts, where digitization and case automation remain limited. It presents a modular, integrated framework tailored to these environments, combining multiple technologies to automate document digitization and support decision-making in HR case processing.

Additionally, this work advances the practical application of machine learning models for imbalanced institutional HR datasets, exploring techniques such as SMOTE, Edited Nearest Neighbors (ENN), and balanced bagging to improve minority class detection. The project demonstrates the use of evaluation metrics aligned with operational priorities, including the F2-Score emphasizing recall, thus providing insights relevant to both academic research and real-world HR system design.

6.3 Limitations of the system

The class skewing present in the institution provisioned dataset resulted in reduced statistical efficiency for the processing of minority class cases. This was, however, counteracted through extensive experimentation to identify the most ideal classification model for the domain context revealed by the data. Additionally, the extraction of data depends largely on the established document structure and text patterns. This then requiring that specific patterns for text extractions must be defined for the handling of different case types. This limitation, further prompting the specialisation of the system in handling only specific case categories before extending to incorporate the handling of other case categories.

6.4 Future works

Future research could explore the further optimisation of the machine learning models to enhance the accuracy of decision making and decision support processes with regards to heavily imbalanced datasets, which appear to be consistent within the domain.

Expanding the system's developmental dataset to incorporate a wider variety of HR cases and larger sample sizes would improve its overall adaptability and add a dynamic robustness to its functionality.

Lastly, the deeper integration of AI decision recommendation with the ABSA module method of data analysis could yield greater benefit in the functional provision of explainable decisions produced within the solution. Rather than the consideration of the AI module as a solution aimed at the replacement of the ABSA functionality, this may perhaps be a beneficial integration that is worthy of exploration towards the functional improvement of the developed solution

References

1. Nawaz, N., Arunachalam, H., Pathi, B. K., & Gajenderan, V. (2024). The adoption of artificial intelligence in human resources management practices. *International Journal of Information Management Data Insights, 4*(1), Article 100208. https://doi.org/10.1016/j.jjimei.2023.100208, El Sevier.
2. Rodgers, W., Murray, J. M., Stefanidis, A., Degbey, W. Y., & Tarba, S. Y. (2023). An artificial intelligence algorithmic approach to ethical decision-making in human resource management processes. *Human Resource Management Review, 33*(1), Article 100925. https://doi.org/10.1016/j.hrmr.2022.100925, El Sevier.
3. Votto, A. M., Valecha, R., Najafirad, P., & Rao, H. R. (2021). Artificial Intelligence in Tactical Human Resource Management: A Systematic Literature Review. *International Journal of Information Management Data Insights, 1*(2), Article 100047. https://doi.org/10.1016/j.jjimei.2021.100047, El Sevier.
4. Kanakov, F., & Prokhorov, I. (2022). Analysis and applicability of artificial intelligence technologies in the field of RPA software robots for automating business processes. *Procedia Computer Science, 213*, 296-300. https://doi.org/10.1016/j.procs.2022.11.070, El Sevier.
5. Górriz, J. M., Álvarez-Illán, I., Álvarez-Marquina, A., Arco, J. E., Atzmueller, M., Ballarini, F., Barakova, E., Bologna, G., Bonomini, P., Castellanos-Dominguez, G., Castillo-Barnes, D., Cho, S. B., Contreras, R., Cuadra, J. M., Domínguez, E., Domínguez-Mateos, F., Duro, R. J., Elizondo, D., Fernández-Caballero, A., Fernandez-Jover, E., ... Ferrández-Vicente, J. M. (2023). Computational approaches to Explainable Artificial Intelligence: Advances in theory, applications and trends. *Information Fusion, 100*, Article 101945, El Sevier. https://doi.org/10.1016/j.inffus.2023.101945.
6. Dafflon, J., Ferreira da Costa, P., Váša, F., & Monti, R. P. (2022). A guided multiverse study of neuroimaging analyses. *Nature Communications, 13*(1), Article 31347. https://doi.org/10.1038/s41467-022-31347-8, Research Gate.
7. S. Wang, M.E. Celebi, Y.-D. Zhang, X. Yu, S. Lu, X. Yao, Q. Zhou, M.-G. Miguel, Y. Tian, J.M. Gorriz, I. Tyukin, Advances in data pre-processing for biomedical data fusion: An overview of the methods, challenges, and prospects, Inf. Fusion 76 (2021) 376–421, http://dx.doi.org/10.1016/j.inffus.2021.07.001.
8. Pandey, M., Arora, M., Arora, S., Goyal, C., Gera, V. K., & Yadav, H. (2023). AI-based integrated approach for the development of intelligent document management system (IDMS). In *3rd International Conference on Evolutionary Computing and Mobile Sustainable Networks (ICECMSN 2023)* (pp. 725–736). https://doi.org/10.1016/j.procs.2023.12.127.
9. J. Memon, M. Sami, R. A. Khan and M. Uddin, "Handwritten Optical Character Recognition (OCR): A Comprehensive Systematic Literature Review (SLR)," in IEEE Access, vol. 8, pp. 142642-142668, 2020, doi: https://10.1109/ACCESS.2020.3012542.
10. Tappert, C. C., Suen, C. Y., & Wakahara, T. (1990). The state of the art in online handwriting recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 12(8), 787–808. doi:10.1109/34.57669.
11. P. Thompson, R. T. Batista-Navarro, G. Kontonatsios, J. Carter, E. Toon, J. McNaught, C. Timmermann, M. Worboys, and S. Ananiadou, ‘‘Text mining the history of medicine,’’ PLoS ONE, vol. 11, no. 1, pp. 1–33, Jan. 2016. doi: https://doi.org/10.1371/journal.pone.0144717.
12. K. D. Ashley and W. Bridewell, ‘‘Emerging AI & Law approaches to automating analysis and retrieval of electronically stored information in discovery proceedings,’’ Artif. Intell. Law, vol. 18, no. 4, pp. 311–320, Dec. 2010, doi: 10.1007/s10506-010-9098-4.

13. R. Zanibbi and D. Blostein, ‘‘Recognition and retrieval of mathematical expressions,’’ Int. J. Document Anal. Recognit., vol. 15, no. 4, pp. 331–357, Dec. 2012, doi: 10.1007/s10032-011-0174-4.
14. S. D. Connell and A. K. Jain, “Template-based online character recognition,” Pattern Recognition, vol. 34, no. 1, pp. 1–14, 2001.
15. A. Graves and J. Schmidhuber, ‘‘Offline handwriting recognition with multidimensional recurrent neural networks,’’ in Proc. Adv. Neural Inf. Process. Syst., 2009, pp. 545–552.
16. Rožanec, J. M., Novalija, I., Zajec, P., Kenda, K., Tavakoli Ghinani, H., & Suh, S. (2022). Human-centric artificial intelligence architecture for industry 5.0 applications. *International Journal of Production Research, 60*(23-24), 6847-6872. https://doi.org/10.1080/00207543.2022.2138611, Taylor and Francis.
17. Palos-Sánchez, P. R., Baena-Luna, P., Badicu, A., & Infante-Moro, J. C. (2022). Artificial Intelligence and Human Resources Management: A Bibliometric Analysis. *Applied Artificial Intelligence: An International Journal, 36*(1), Article 2145631. https://doi.org/10.1080/08839514.2022.2145631, Taylor and Francis.
18. Russell, S., & Norvig, P. (2010). *Artificial intelligence: A modern approach* (3rd ed.). Prentice Hall.
19. Negnevitsky, M. (2005). *Artificial intelligence: A guide to intelligent systems* (2nd ed.). Pearson Education.
20. Vrontis, D., Christofi, M., Pereira, V., Tarba, S., Makrides, A., & Trichina, E. (2021). Artificial intelligence, robotics, advanced technologies and human resource management: a systematic review. The International Journal of Human Resource Management, 1–30. doi:10.1080/09585192.2020.1871398, Taylor and Francis.
21. Parry, E., & Tyson, S. (2008). An analysis of the use and success of online recruitment methods in the UK. Human Resource Management Journal, 18(3), 257–274. https://doi.org/10.1111/j.1748-8583.2008.00070.x
22. Bhave, D. P., Teo, L. H., & Dalal, R. S. (2020). Privacy at work: A review and a research agenda for a contested terrain. Journal of Management, 46(1), 127–164. https://doi.org/10.1177/0149206319878254.
23. Stone, D. L., Deadrick, D. L., Lukaszewski, K. M., & Johnson, R. (2015). The influence of technology on the future of human resource management. Human Resource Management Review, 25(2), 216–231. https://doi.org/10.1016/j.hrmr.2015.01.002.
24. Bondarouk, T., Parry, E., & Furtmueller, E. (2017). Electronic HRM: Four decades of research on adoption and consequences. The International Journal of Human Resource Management, 28(1), 98–131. https://doi.org/10.1080/09585192.2016.1245672.
25. Cooke, F. L., Wood, G., Wang, M., & Veen, A. (2019). How far has international HRM travelled? A systematic review of literature on multinational corporations (2000–2014). Human Resource Management Review, 29(1), 59–75. https://doi.org/10.1016/j.hrmr.2018.05.001.
26. Abraham, M., Niessen, C., Schnabel, C., Lorek, K., Grimm, V., Moslein, K., & Wrede, M. (2019). Electronic monitoring at work: The role of attitudes, functions, and perceived control for the acceptance of tracking technologies. Human Resource Management Journal, 29(4), 657–675. https://doi.org/10.1111/1748-8583.12250.
27. Parry, E., & Tyson, S. (2011). Desired goals and actual outcomes of e-HRM. Human Resource Management Journal, 21(3), 335–354. https://doi.org/10.1111/j.1748-8583.2010.00149.x

28. Grace, K., Salvatier, J., Dafoe, A., Zhang, B., & Evans, O. (2018).Viewpoint: When will AI exceed human performance? Evidence from AI experts. Journal of Artificial Intelligence Research, 62,729–754. https://doi.org/10.1613/jair.1.11222
29. Ozmen Garibay, O., Winslow, B., Andolina, S., Antona, M., Bodenschatz, A., & Coursaris, C. (2023). Six Human-Centered Artificial Intelligence Grand Challenges. *Journal of Management Information Systems*, 40(2), 391-437. https://doi.org/10.1080/10447318.2022.2153320
30. Xu, Z. (2021). Human Judges in the Era of Artificial Intelligence: Challenges and Opportunities. *Applied Artificial Intelligence: An International Journal*, Advance online publication. https://doi.org/10.1080/08839514.2021.2013652
31. Volokh, E. 2019. Chief Justice Robots. Duke Law Journal 68(6):1134–92
32. Manneh, Y. E., & Adesopo, A. (2022). Effect of recruitment and selection methods on employee performance in the public service of the Gambia. *Canadian Social Science, 18*(1), 109-123.
33. Choudrie, J., & Dwivedi, Y. K. (2005). Investigating the research approaches for examining technology adoption issues. *Journal of Research Practice, 1*(1), 1-12.
34. Bagdasarov, Z., Martin, A. A., & Buckley, M. R. (2020). Working with robots: Organizational considerations. *Organizational Dynamics, 49*(2), 100679.
35. Melesse, T. Y., Di Pasquale, V., & Riemma, S. (2020). Digital twin models in industrial operations: A systematic literature review. *Procedia Manufacturing, 42*, 267-272.
36. Li, P., Bastone, A., Mohamad, T. A., & Schiavone, F. (2023). How does artificial intelligence impact human resources performance? Evidence from a healthcare institution in the United Arab Emirates. *Journal of Innovation & Knowledge, 8*(2), 100340. https://doi.org/10.1016/j.jik.2023.100340.
37. Vrontis, D., Christofi, M., Pereira, V., Tarba, S., Makrides, A., & Trichina, E. (2022). Artificial intelligence, robotics, advanced technologies, and human resource management: A systematic review. *International Journal of Human Resource Management, 33*(6), 1237-1266. https://doi.org/10.1080/09585192.2020.1871398.
38. Li, J. (Justin), Bonn, M. A., & Ye, B. H. (2019). Hotel employees' artificial intelligence and robotics awareness and its impact on turnover intention: The moderating roles of perceived organizational support and competitive psychological climate. *Tourism Management, 73*(December 2018), 172-181. https://doi.org/10.1016/j.tourman.2019.02.006.
39. Bhardwaj, G., Singh, S. V., & Kumar, V. (2020). An empirical study of artificial intelligence and its impact on human resource functions. In *Proceedings of the International Conference on Computing, Automation, and Knowledge Management* (ICCAKM), 47-51. https://doi.org/10.1109/ICCAKM46823.2020.9051544.
40. Chatterjee, S., Rana, N. P., Khorana, S., Mikalef, P., & Sharma, A. (2021). Assessing organizational users' intentions and behavior to AI-integrated CRM systems: A meta-UTAUT approach. *Information Systems Frontiers*. https://doi.org/10.1007/s10796-021-10181-1.
41. Wang, S. M., Huang, Y. K., & Wang, C. C. (2020). A model of consumer perception and behavioral intention for AI service. In *Proceedings of the ACM International Conference*, 196-201. https://doi.org/10.1145/3396743.3396791.
42. Butler, M. G., & Callahan, C. M. (2014). Human resource outsourcing: Market and operating performance effects of administrative HR functions. *Journal of Business Research, 67*(2), 218-224. https://doi.org/10.1016/j.jbusres.2012.09.026.

43. Santana, M., & Díaz-Fernández, M. (2022). Competencies for the Artificial Intelligence Age: Visualisation of the State of the Art and Future Perspectives. *Springer Berlin Heidelberg*. https://doi.org/10.1007/s11846-022-00613-w.
44. Hamid, H., & Zeki, A. M. (2014). Users' awareness of and perception on information security issues: A case study of kulliyyah of ICT postgraduate students. In *Proceedings of the 3rd International Conference on Advanced Computer Science Applications and Technologies* (ACSAT), 139-144. https://doi.org/10.1109/ACSAT.2014.31.
45. Brougham, D., & Haar, J. (2018). Smart technology, artificial intelligence, robotics, and algorithms (STARA): Employees' perceptions of our future workplace. *Journal of Management & Organization, 24*(2), 239-257. https://doi.org/10.1017/jmo.2016.55.
46. Shahzad, M. F., Xu, S., Naveed, W., Nusrat, S., & Zahid, I. (2023). Investigating the impact of artificial intelligence on human resource functions in the health sector of China: A mediated moderation model. *Heliyon, 9*, e21818. https://doi.org/10.1016/j.heliyon.2023.e21818.
47. Brewster, C., & Hegewisch, A. (1994). Human resource management in Europe: Issues and opportunities. In *Policy and practice in European human resource management: The Price Waterhouse Cranfield Survey*.
48. Jatobá, M., Santos, J., Gutierrez, I., Moscon, D., Fernandes, P. O., & Teixeira, J. P. (2019). Evolution of artificial intelligence research in human resources. *Procedia Computer Science, 164*, 137-142. https://doi.org/10.1016/j.procs.2019.12.165.
49. Nascimento, M. A., & Queiroz, M. C. A. (2017). Overview of research on Artificial Intelligence in Administration in Brazil. In *ANPAD Meetings – Enanpad, 2017* (pp. 1-4).
50. Ruby Merlin.P1, Jayam.R2, Artificial Intelligence in Human Resource Management, International Journal of Pure and Applied Mathematics, Volume 119 No. 17 2018,1891-1895 ISSN: 1314-3395
51. F. David Schoorman, "Escalation Bias in Performance Appraisals: An Unintended Consequence of Supervisor Participation in Hiring Decisions," Journal of Applied Psychology, 73/1 (1988): 58-62.
52. Tambe, P., Cappelli, P., & Yakubovich, V. (2019). *Artificial Intelligence in Human Resources Management: Challenges and a Path Forward. California Management Review, 61(4), 15–42.* doi:10.1177/0008125619867910.
53. Cappelli, P., "There's No Such Thing as Big Data in HR," Harvard Business Review Digital Articles, June 2, 2017, pp. 2-4.
54. Junque de Fortune, E., Martens, D., & Provost, F., "Predictive Modeling with Big Data: Is Bigger Really Better?" Big Data, 1/4 (December 13, 2004): 215-226.
55. Lee, M.K., Kusbit, D., Metsky, E., & Dabbish, L.A., "Working with Machines: The Impact of Algorithmic and Data-Driven Management on Human Workers," in Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (New York, NY: ACM, April 2015), pp. 1603-1612.
56. Netessine, S., & Yakubovich, V., "The Darwinian Workplace," Harvard Business Review, 90/5 (May 2012): 25-28.
57. European Union. (n.d.). *EU General Data Protection Regulation (GDPR)*. Retrieved November 2, 2024, from https://www.eugdpr.org.
58. Areheart, B., & Roberts, J., "GINA, Big Data, and the Future of Employee Privacy," Yale Law Journal, 128/3 (2019): 710-790.

59. Dwork, C., & Roth, A., The Algorithmic Foundations of Differential Privacy (Boston, MA: Now Publishers, 2014), p. 5.
60. Daugherty, P. R., & Wilson, H. J. (2018). *Human + machine: Reimagining work in the age of AI*. Harvard Business Press
61. Ågerfalk, P. J. (2020). Artificial intelligence as digital agency. *European Journal of Information Systems*, *29*(1), 1–8. https://doi.org/10.1080/0960085X.2020.1721947
62. Bughin, J., Batra, P., Chui, M., Manyika, J., Ko, R., Sanghvi, S., Woetzel, J., & Lund, S. (2017). *Jobs lost, jobs gained: Workforce transitions in a time of automation*. McKinsey Global Institute.
63. Osterlund, C., Jarrahi, M. H., Willis, M., Boyd, K., & Wolf, C. T. (2021). Artificial intelligence and the world of work, a co-constitutive relationship. *Journal of the Association for Information Science and Technology*, *72*(1), 128–135. https://doi.org/10.1002/asi.24388.
64. Liu, B. (2012). Sentiment analysis and opinion mining. Synthesis Lectures on Human Language Technologies, 5(1), 1–167. doi:10.2200/S00416ED1V01Y201204HLT016.
65. Pang, B., & Lee, L. (2006). Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval, 1(2), 91–231. doi:10.1561/1500000001.
66. Mowlaei, M. E., Abadeh, M. S., & Keshavarz, H. (2020). *Aspect Based Sentiment Analysis using Adaptive Aspect Based Lexicons. Expert Systems with Applications, 113234.* doi:10.1016/j.eswa.2020.113234
67. Keshavarz, H., & Abadeh, M. S. (2016). SubLex: Generating subjectivity lexicons using genetic algorithm for subjectivity classification of big social data. In 1st Conference on Swarm Intelligence and Evolutionary Computation, CSIEC 2016 - Proceedings. doi:10.1109/CSIEC.2016.7482126.
68. Akhtar, M. S., Gupta, D., Ekbal, A., & Bhattacharyya, P. (2017). Feature selection and ensemble construction: A two-step method for Aspect Based sentiment analysis. Knowledge-Based Systems, 125, 116–135. doi:10.1016/j.knosys.2017.03.020.
69. Awwad, H., & Alpkocak, A. (2016). Performance comparison of different lexicons for sentiment analysis in Arabic. In Proceedings - 2016 3rd European Network Intelligence Conference, ENIC 2016. doi:10.1109/ENIC.2016.026.
70. Han, H., Zhang, J., Yang, J., Shen, Y., & Zhang, Y. (2018). Generate domain-specific sentiment lexicon for review sentiment analysis. Multimedia Tools and Applications, 1–16. doi:10.1007/s11042-017-5529-5.
71. Keshavarz, H., & Abadeh, M. S. (2017). ALGA: Adaptive lexicon learning using genetic algorithm for sentiment analysis of microblogs. Knowledge-Based Systems, 122, 1–16. doi:10.1016/j.knosys.2017.01.028.
72. Hoang, M., Bihorac, O. A., & Rouces, J. (2019). Aspect Based sentiment analysis using BERT. *In Proceedings of the 22nd Nordic Conference on Computational Linguistics* (pp. 187–196). Linköping University Electronic Press. https://www.aclweb.org/anthology/W19-5317.
73. Pontiki, M., Galanis, D., Pavlopoulos, J., Papageorgiou, H., Androutsopoulos, I., & Manandhar, S. (2014). SemEval-2014 task 4: Aspect Based sentiment analysis. *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, 27–35.
74. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational*

Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 4171–4186. Association for Computational Linguistics.
75. Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 2227–2237. Association for Computational Linguistics.
76. Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Kaiser, Ł., Gouws, S., Kato, Y., Kudo, T., Kazawa, H., Stevens, K., Kurian, G., Patil, N., Wang, W., Young, C., Smith, J., Riesa, J., Rudnick, A., Vinyals, O., Corrado, G., Hughes, M., & Dean, J. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv*. https://arxiv.org/abs/1609.08144
77. Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv*. https://arxiv.org/abs/1301.3781.
78. Pontiki, M., Galanis, D., Papageorgiou, H., Androutsopoulos, I., Manandhar, S., Al-Smadi, M., Al-Ayyoub, M., Zhao, Y., Qin, B., De Clercq, O., Hoste, V., Apidianaki, M., Tannier, X., Loukachevitch, N., Kotelnikov, E., Bel, N., Jiménez-Zafra, S. M., & Eryiğit, G. (2016). SemEval-2016 Task 5: Aspect Based sentiment analysis. *In Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)* (pp. 19–30). Association for Computational Linguistics.
79. Che, W., Zhao, Y., Guo, H., Su, Z., & Liu, T. (2015). Sentence compression for Aspect Based sentiment analysis. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 23*(12), 2111–2124. https://doi.org/10.1109/taslp.2015.2443982.
80. Taboada, M., Brooke, J., Tofiloski, M., Voll, K., Stede, M., Lexicon-based methods for sentiment analysis, Comput. Linguist. 37 (2) (2011) pp. 267–307.
81. AIMultiple. (2024). *OCR accuracy: Measuring the accuracy of OCR solutions*. Retrieved December 30, 2024, from https://research.aimultiple.com/ocr-accuracy/.
82. Rakshit, S., & Basu, S. (2009). Development of a multi-user handwriting recognition system using Tesseract open source OCR engine. *Proceedings of the International Conference on C3IT (2009)*, 240–247. https://arxiv.org/abs/1003.5886.
83. Hmoud, B., & Várallyai, L. (2019). Will artificial intelligence take over human resources recruitment and selection? *Network Intelligence Studies, 7*(13), 21, from https://www.researchgate.net/publication/337931190.
84. Consoli, S., Barbaglia, L., & Manzan, S. (2022). Fine-grained, aspect-based sentiment analysis on economic and financial lexicon. *Knowledge-Based Systems, 246*, 108781. https://doi.org/10.1016/j.knosys.2022.108781.
85. D. J. Burr, "Designing a handwriting reader," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-5, no. 5, pp. 554–559, 1983, doi: 10.1109/TPAMI.1983.4767435.
86. B. P. Berman and R. J. Fateman, "Optical character recognition for typeset mathematics," in *Proc. ACM Conf. Document Processing Systems*, Jan. 1994, doi: 10.1145/190347.190438.
87. U. Garain, "Identification of mathematical expressions in document images," in *Proc. Int. Conf. Document Analysis and Recognition*, Barcelona, 2009, pp. 1340–1344. doi: 10.1109/ICDAR.2009.203.

88. A. Graves and J. Schmidhuber, “Offline handwriting recognition with multidimensional recurrent neural networks,” in *Proc. 22nd Int. Conf. Neural Information Processing Systems (NIPS)*, 2008, pp. 545–552.
89. C. Malon, S. Uchida, and M. Suzuki, “Mathematical symbol recognition with support vector machines,” *Pattern Recognition Letters*, vol. 29, no. 9, pp. 1326–1332, 2008, doi: 10.1016/j.patrec.2008.02.005.
90. N. A. Jebril, H. R. Al-Zoubi, and Q. A. Al-Haija, “Recognition of handwritten Arabic characters using histograms of oriented gradient (HOG),” *Pattern Recognition Image Analysis*, vol. 28, no. 2, pp. 321–345, 2018.
91. Lenz Furrer and Martin Volk, “Reducing OCR errors in gothic-script documents,” in *Proc. Workshop on Language Technologies for Digital Humanities and Cultural Heritage*, 2011, pp. 97–103.
92. A. Malik, S. NR, and P. Budhwar, “Digitisation, Artificial Intelligence (AI) and HRM,” in *Human Resource Management: Strategic and International Perspectives*, J. Crawshaw, P. Budhwar, and A. Davis, Eds., 3rd ed. London, U.K.: SAGE Publications, 2020, pp. 88–111.
93. Carl Lawrence, Tuure Tuunanen, and Michael D. Myers, "Extending Design Science Research Methodology for a Multicultural World," in *Human Benefit through the Diffusion of Information Systems Design Science Research*, J. Pries-Heje, J. R. Venable, D. Bunker, N. L. Russo, and J. I. DeGross, Eds., Perth, Australia: Springer, IFIP Adv. Inf. Commun. Technol., vol. 318, pp. 108–121, 2010. doi: 10.1007/978-3-642-12113-5_9.
94. T. M. Takkimsncn, *HR Classification for Promotion* [Online]. Available: https://www.kaggle.com/code/takkimsncn/hr-classification-for-promotion/input.
95. A. Nic, HR Analytics: Employee Promotion Data – Predict the eligible candidates for promotion [Online]. Available: https://www.kaggle.com/datasets/arashnic/hr-ana [Accessed: 06-March-2025].
96. R. Tulluri, *Employee Promotion Prediction Dataset* [Online]. Available: https://github.com/rajtulluri/EmployeePromotionPrediction/blob/master/employeePromotion.csv [Accessed: 06-March-2025].
97. OpenCV, *Image Thresholding — OpenCV-Python Tutorials 1 documentation*, [Online]. Available: https://docs.opencv.org/4.x/d7/d4d/tutorial_py_thresholding.html. [Accessed: 06-March-2025].

APPENDIX

Appendix I : System User Interface