National College of Ireland

NLP and Machine Learning Framework for Reducing False Positive in Watchlist Name Screening in AML Systems

MSc Research Project
MSc Data Analytics

Forename Surname
Student ID: 23407476

School of Computing
National College of Ireland

Supervisor: Hamilton Niculescu

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Gayatri Jadhav
Student ID:	23407476
Programme:	MSc Data Analytics **Year:** 2025-26
Module:	MSc Research Practicum
Supervisor:	Hamilton Niculescu
Submission Due Date:	28/01/2026
Project Title:	NLP and Machine Learning Framework for Reducing False Positive in Watchlist Name Screening in AML Systems
Word Count:	8377 **Page Count:** 20

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Gayatri Jadhav
Date:	23/01/2026

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Abstract

The Anti-Money laundering (AML) systems screening sanctions often give high false-positive rates as a result of linguistic variation, lack of consistency in transliteration and typing errors. Conventional rule-based and fuzzy string-matching algorithms, e.g. Levenshtein and Jaro-Winkler, typically have a hard time representing semantic relationships between names, leading to higher operational costs and regulatory inefficiencies.

This study created a hybrid Natural Language Processing (NLP) and Machine Learning (ML) model that combines Sentence-BERT (SBERT) semantic embeddings with traditional string-similarity scores with the help of Shapley Additive Explanations (SHAP) to make the results easier to interpret. Synthetic customer records were generated using standard data-generation techniques to maintain privacy. Random Forest, XGBoost and Support Vector Machine (SVM) which are three supervised classifiers were trained on the integrated semantic-fuzzy feature set and assessed in terms of Precision, Recall, F1-score, ROC-AUC and False Positive Rate.

The discrimination of matching and non-matching name pairs was moderate in all models. SVM had the best generalisation with the highest ROC-AUC score, and the other two models, Random Forest and XGBoost, had more recall at the expense of more false positives. SHAP analysis revealed that fuzzy similarity measures, especially TokenSetRatio, had the strongest impact on all models, and the contribution of features was in line with human intuition.

In theory, the findings demonstrate that contextual embeddings can be used to supplement surface-level similarity in short-text name-matching tasks. In practice, the hybrid framework will increase the transparency, interpretability and auditability of AML screening processes, which is in line with regulatory requirements of explainable AI. More advancements in predictive accuracy will need more multilingual and phonetic and relational features and access to real operational data.

Chapter 1: Introduction

1.1 Background

Money laundering, terrorism financing, and evasion of sanctions are constant threats to the global financial system. Financial institutions must go to the extent of enacting Anti-Money Laundering (AML) and sanctions screening policies to meet the international regulatory standards such as those mandated by Financial Action Task Force (FATF), the United Nations (UN), the European Union (EU) and the U.S. Office of Foreign Assets Control (OFAC) to counter such threats. Sanctions screening is part of the AML compliance, involves identifying individuals, organisations or entities that cannot engage in financial transactions, because they are connected with criminal or terrorist activities.

Traditional techniques of sanctions screening are based mostly on rule-based or fuzzy string-matching algorithms like Levenshtein or Jaro-Winkler distances (Pikies and Ali, 2021). Although the methods have worked in the previous stages of AML implementation, they struggle to cope with the complexity of modern, globalised financial environments. These legacy systems perform poorly when confronted with linguistic diversity, transliteration inaccuracies, typos and language naming conventions found in customer databases. As a result, they generate disproportionately high false-positive rates, creating substantial operational burdens and increasing compliance costs due to extensive manual review.

Artificial Intelligence and Natural Language Processing have become increasingly influential technologies in financial compliance in the past years (Kalusivalingam, et al., 2022). The transformer-based models like Sentence-BERT have proven their capability to capture the semantic similarity of names, which makes them more context-sensitive and more accurate in the screening results. Modern hybrid algorithms can provide a future direction in reducing false positives and making screening systems more explainable by using semantic embeddings with classical string similarity measures.

This study builds upon these technological advancements by proposing a hybrid, explainable AI framework for sanctions name screening. The integration of semantic similarity, fuzzy matching, and explainability tools such as Shapley Additive Explanations (SHAP) aims to improve operational efficiency, maintain regulatory transparency, and address longstanding challenges in AML screening systems.

1.2 Problem Domain

Even with the development of AI and ML, sanctions screening still has significant difficulties with accuracy, scalability, and transparency of compliance (Hu, 2024). The most critical and longstanding issue is the high volume of false positives generated by existing systems. The slightest difference in spelling, transliteration mistakes or alias may cause legitimate customer transactions to be flagged by financial institutions. Such false alarms not only consume resources, but they also destroy customer relations and delay legitimate transactions.

A second major challenge is that of data integration and standardisation. International sanctions lists issued by bodies such as the UN, EU, and OFAC differ in structure, format, update frequency, and naming conventions. This contradiction makes it difficult to have a consolidated and effective screening database. Furthermore, the central databases of most financial institutions, particularly in developing countries, do not have a sound data governance infrastructure, resulting in unfinished or uneven Know Your Customer (KYC) information, which also reduces the accuracy of the screening.

Moreover, the transition to AI-oriented models creates another dimension of challenges, namely, explainability and regulatory trust (De Almeida et al., 2021). Regulators not only expect financial institutions to identify risks correctly but also to justify automated decisions. Although models like SBERT that are based on a transformer can be very high-performing, their black-box nature restricts transparency. This lack of transparency presents a significant barrier to adoption within highly regulated financial sectors.

The literature suggests that a majority of AML research have concentrated on shallow machine learning or fuzzy-matching schemes, and there has been limited literature regarding contextual embeddings and explainable hybrid models. This creates a gap in developing high-performing, interpretable, and multilingual AML screening systems (Ghimire, 2025). To address these issues, the proposed research will use semantic and syntactic methods supported by SHAP-based interpretability aiming to provide an accurate, transparent, and regulator-aligned name-screening framework for AML systems.

1.3 Aim and Objectives

Aim
The aim of this research is to create a hybrid NLP and machine learning-based model that can successfully minimize false positives in AML sanctions name-screening while enhancing model transparency and regulatory explainability.

Research Questions
To what extent can a hybrid NLP–ML framework reduce false positives while enhancing transparency in AML sanctions screening?

Objectives
To achieve the research aim, the study is guided by the following objectives:

1. **To critically examine the limitations** of rule-based and fuzzy-matching algorithms in sanctions name screening, especially with regard to false positives, transliteration and language variability.
2. **To design and execute a hybrid system** that integrates Sentence-BERT semantic embeddings with the standard string distance metrics such as Levenshtein and Jaro-Winkler.
3. **To train and compare supervised machine learning classifiers**, such as XGBoost, Random Forest, and Support Vector Machine (SVM), on the basis of their performance in terms of minimizing false positives and high recall in true matches.
4. **To measure the model performance** according to the typical metrics of Precision, Recall, F1-Score, ROC-AUC, and False Positive Rate (FPR), achieving the balance between predictive performance and compliance efficiency.

These objectives will enable the research to make a new, interpretable, and cost-efficient solution to AML name screening systems that is in line with the global compliance requirements and frailties of the current solution in operations.

1.4 Rationale

The arguments informing this research lie in the increasing need for accuracy, efficiency, and transparency in the AML sanctions screening processes. Existing compliance systems face persistent challenges in balancing strict regulatory expectations with operational practicality. Traditional methods generate high false-positive which in addition to increasing the operational costs, and negatively impact customer experience. This study provides a workable and scalable solution to these problems that plague the industry by incorporating contextual NLP embeddings, and explainable AI mechanisms (Zini and Awad, 2022).

With advancements in NLP and AI, syntactic similarity and semantic similarity hybridization is a new step. Semantic models such as SBERT read between the lines and can see the contextual associations as well as linguistic peculiarities, whereas string similarity algorithms deal with typographical and orthographic anomalies (Ndimbo, et al., 2025). Combination of these complementary perspectives enables the system to obtain meaning as well as the structure of names hence enhancing accuracy and minimizing misclassification. Equally important is the integration of Explainable AI (XAI). SHAP offers detailed insights into model predictions, allowing compliance officers and auditors to understand why a name was flagged. This aspect makes AI applications in the financial sector more credible and accountable as per the global expectations for accountable AI.

Overall, this study is ethically oriented since it makes use of synthetic data generation, thus, preventing the risk of infringing the privacy of the customers or the data protection laws (Shehu and Shehu, 2023). This approach does not only ensure that the bias is reduced, but also makes the strategy inclusive to reflect various conventions of names between different cultures and languages, which is a crucial aspect of creating fair AI systems to support a globalized economy. To conclude, this study is practically and academically important. It adds to the literature in the sense that it helps to fill the gap between AML compliance and more sophisticated AI modelling, and also provides financial institutions with a feasible way of more precise, transparent, and resource-efficient sanctions screening.

Chapter 2: Literature Review

2.1 Existing Research on Sanctions Screening and AML Compliance

The current compliance procedures of the AML suggests that the sanction screening is just a small proportion of the bigger steps that are taken to identify and combat the illegal financial activity. It has been even verified that the screening has been carried out as a real time checking of the data of the customers against the existing lists of sanctions issued by the various sanction authorities which are the United Nations (UN), the United States Office of Foreign Assets Control (OFAC), the European Union (EU) and with other authorities (Ciapriani 2023).

This expansion of screening responsibilities shifts the duties of financial institutions towards a more active obligation since screening now falls between risk profiling, as part of the Customer Due Diligence (CDD) process, and transaction monitoring. This overlap can lead to ambiguity regarding appropriate level of scrutiny required at different stages of the compliance process.
In the existing literature, NLP and fuzzy matching algorithms have been investigated as solutions to the challenges inherent in name variation, transliteration, and misspellings in sanctions data (Wang, 2022). Regardless of technological advancements, it has been researched that ML models can provide a basis for prioritising alerts, producing cost savings, and not undermining the rigorous approach required in regulatory frameworks.

Research also recognises the tension between compliance effectiveness and operational efficiency. High false-positive volumes increase manual workload, drive up operational costs, and may delay legitimate customer transactions. In parallel, global disparities in technology adoption further complicate compliance. While financial institutions in developed jurisdictions have begun incorporating advanced AML technologies, many emerging economies still rely heavily on manual processes or outdated systems. This creates uneven enforcement and increases vulnerability to regulatory arbitrage, where illicit actors exploit weakly supervised jurisdictions.

Overall, the literature demonstrates that the process of screening sanctions is not necessarily a technical issue but a regulatory one and also an organisational problem. Johnson (2025) infers that anti-money laundering compliance in the future will stabilise regulatory demands, act responsibly with AI technologies, and cooperate internationally, at least, taking into consideration the standards of enforcing regulations in various jurisdictions.

2.2 Review of Techniques for Sanctions Screening in Anti-Money Laundering

The development of sanctions screening methods has a history of evolving from older techniques comprised of rule-based screening, to specialised sanctions screening methods based on computational, fuzzy logic and AI techniques.
Historically, exact matching techniques were widely used, but these approaches were unable to handle the complexities of name variations, transliterations, abbreviations, and cultural naming differences.
The difficulties associated with processing name variations remained because screening of sanctions was transferred into a global context, and incomplete methods based on rules resulted in high false positives. Increasingly complex fuzzy matching algorithms started to supplant the simple ones.

To address these limitations, fuzzy-matching algorithms and phonetic techniques were introduced. Among the algorithms that are frequently mentioned in the literature as able to identify some characters without matching them are Soundex, Levenshtein distance, and Jaro-Winkler (Afiyati et al., 2023). These procedures became more accurate, but could not solve the issue of multi-lingual data processing (weakness in processing large volumes of data), and research shifted to hybrid schemes involving deterministic rules combined with probabilistic models. Machine learning and the application of AI in sanctions screening have also been welcomed into the field. Supervised learning algorithms such as decision trees and support vector machines, which have been used to classify alerts and take high-risk entities into account.

Parallel developments in NLP have opened new possibilities. Transformer-based architectures, including BERT and its variants, have demonstrated exceptional performance in semantic similarity tasks and entity matching. Convolutional Neural Nets (CNNs) and Recurrent Neural Nets (RNNs) are also being recognised and used to do name-matching, and have been reported to achieve improvements in their analysis.

2.3 International Perspectives on Sanctions Data Management in AML

The further development of the internationalisation of financial regimes has given rise to the increased attention paid to the sanctions data governance as the basic component of the AML applications. The literature highlights on the uneven implementation of sanctions management practices in differing jurisdictions. While developed countries are deploying robust automated systems, developing economies are often using ineffective, antiquated systems with inadequate

regulatory regimes. Research papers have identified that the absence of operational systems to exchange information creates gaps that illegal actors exploit to evade jurisdiction (Voss, 2021).

The analysis of blockchain-based solutions to manage sanctions lists, where potentially updates could be made immutable, or decentralised, accessible everywhere in the world. Similarly, the work of Jain (2021) cloud-based regtech platforms are being advertised as scalable to a larger number of smaller financial institutions in developing economies. Nevertheless, there are issues with cybersecurity, governance, and interoperability with these innovations. In addition, the political considerations of sanctions are noted in the global outlook. Sanctions are more than just compliance; they are also tools of foreign policy. This two-sided nature makes it difficult to implement and manage AML, as the institution has two competing national political interests to balance while remaining neutral.
Murrar and Barakat (2021) have discussed that the international co-operation of countries in programs developed by organisations such as the FATF and IMF is often directed toward increasing the capacity of countries to implement sanctions properly.

2.4 Contextual Embeddings and Explainable AI in AML Systems (Transformer-Based)

2.4.1 Entity Matching and Transformer-Based Language Models

Ditto is an entity-matching system that uses pre-trained transformer models, such as BERT and RoBERTa, to determine whether two textual records refer to the same entity (Li et al., 2020).The experiment model redefined record matching to a sequence-pair classification problem, which enabled the model to learn the semantic dependencies that could not be identified using the traditional string matching or fuzzy-matching algorithms.

Experimental results demonstrate that transformer-based embeddings outperform past heuristic and machine-learning-based methods on numerous benchmark datasets. Transformer-based models show strong resilience to linguistic variation, transliteration differences, and domain-specific terminology. It would imply, in the context of sanction screening systems, that BERT-based models would be in a position to be more reasonable when identifying a sanctioned entity, as either the name has been shortened, transliterated, or masqueraded as an alias, which, in most cases of traditional systems, would lead to a false positive notification.

2.4.2 Relation-Aware Entity Matching Using Sentence-BERT

Considering the paradigm of transformers, Zhou et al. (2022) designed a Relation-Aware Sentence-BERT model so that the problem of semantic entity-matching could be addressed. Sentence-BERT was improved on by introducing relational and contextual information between pairs of entities as compared to the case of treating entities in individual sentences. According to the experimental findings, it was better at the name-matching and semantic-similarity tasks compared to the baseline models of standard SBERT and BioBERT. The specified method can be utilised particularly well in the framework of sanction screening, in which case the relationship between organisations, along with their aliases, subsidiaries, or other individuals connected to them, can predetermine the compliance outcome.

2.4.3 SHAP-Based Model Interpretability of Financial Systems

The use of XAI in financial modelling, which was introduced by Lin and Gao (2022) with SHAP added to the fraud-detection processes, has also become a part of the model usage. Their Group SHAP method aggregates feature contributions across predictions, enabling clear and comprehensive visualisations while preserving predictive accuracy. The paper emphasised the role of openness in high-stakes financial situations that claiming that the black-box models discourage regulatory trust.
The authors have demonstrated that SHAP explanations allowed the compliance analysts to identify the best and most robust behavioural and transactional attributes that lead to flagged alerts (Lin and Gao, 2022). The article provides sound empirical evidence that SHAP can afford to place the transparency demands of the modern compliance regimes.

2.5 Research Gap Analysis

Theme / Area	Existing Literature Findings	Identified Research Gap	Relevance to Current Study
Sanctions Screening and AML Compliance	Most studies (Cipriani et al., 2023; Johnson, 2025) focus on rule-based or fuzzy-matching methods for sanctions name screening. These approaches improve efficiency but still yield high false positives.	Limited exploration of deep contextual models (e.g., Sentence-BERT) in sanctions screening; lack of integrated semantic understanding for multilingual or alias-heavy data.	The current study applies transformer-based contextual embeddings (Sentence-BERT) to improve accuracy and reduce false-positive rates in sanctions name-matching.
Data Integration and Standardisation	Literature (Azeroual et al., 2023; Rahim et al., 2024) discusses challenges in data harmonisation and governance across global sanctions lists.	Existing work does not integrate AI-driven entity resolution with explainable compliance frameworks.	The proposed framework enhances data standardisation using contextual embeddings while maintaining transparency through SHAP interpretability.
Fuzzy Matching and NLP in Sanctions Systems	Research (Wang, 2022; Afiyati et al., 2023) explores fuzzy matching and NLP for name variation issues.	Focus remains on surface-level string similarity; limited attention to contextual semantics or multi-language adaptability.	This research introduces transformer-based semantic matching using Sentence-BERT, capable of capturing deeper contextual and linguistic relationships.
Explainable AI (XAI) in Financial Compliance	Studies (Lin and Gao, 2022; Zhang et al., 2022) apply SHAP to fraud detection for model interpretability.	No prior integration of SHAP within sanctions screening frameworks; explainability in sanction models remains underdeveloped.	The study introduces SHAP-based explanation layers to interpret Sentence-BERT-driven AML models, improving transparency and regulatory acceptance.
Integration of Deep Learning with Regulatory Context	Existing works either focus on detection performance (Li et al., 2020; Zhou et al., 2022) or interpretability (Lin and Gao, 2022), but not both.	Absence of hybrid frameworks that combine high-performance deep learning (BERT/SBERT) with explainable AI for AML.	This study fills the integration gap by proposing a hybrid, interpretable AI model that leverages Sentence-BERT for detection and SHAP for explanation.
Cross-Jurisdictional AML Compliance Challenges	Global studies (Murrar and Barakat, 2021; Voss, 2021) highlight uneven enforcement and inconsistent data-sharing frameworks.	Lack of adaptable, transparent AI models that can operate across regulatory jurisdictions.	The proposed model demonstrates generalisability and interpretability to support multi-jurisdictional AML compliance.

Following this analysis, it is evident that existing research does not adequately address the need for a semantically aware, interpretable, and regulator-friendly sanctions-screening model. This study therefore aims to bridge these gaps by proposing a hybrid AML framework that integrates Sentence-BERT embeddings for contextual understanding with classical string-similarity metrics for surface accuracy, supported by SHAP-based explainability to meet regulatory requirements. The resulting system seeks to provide a more efficient, transparent, and operationally meaningful approach to AML name screening.

Chapter 3: Research Methodology

The paper is set to formulate a hybrid NLP model and machine learning (ML) model, which can be used to reduce false positives during the use of the watchlist name screening in Anti-Money Laundering (AML). The traditional AML systems are rule-based or fuzzy string-matching, e.g., the Levenshtein distances or the Jaro-winkler distances, and are limited in their capability of addressing changes in name, phonetic cases, transliteration and typing mistakes. As a result, these systems are prone to generating excessive false alarms and generating excessive costs in manual countercheck and operational wastages (Gandhi et al., 2024).

The technique used to address these problems by using Sentence-BERT embeddings, traditional string similarity capabilities, and supervised machine deepens is described in this chapter and justified with the help of SHAP. It mentions

the plans of research design, data collection and preprocessing, feature engineering, model development, assessment plans, and ethics.

3.1 Research Design

The study follows a quantitative, experimental type of research, which involves machine learning modelling and NLP. It uses a hybrid method of combining semantic similarity from Sentence-BERT embeddings with classical string similarity.

This design is motivated by three key considerations:

1. **Accuracy Enrichment:** Semantic embeddings are able to take into account the contextual and phonetic context of names that fuzzy matching cannot.
2. **Explainability and Compliance:** SHAP offers interpretable model results, which will allow the financial institutions to audit the decision-making in accordance with the regulations (Alkhalili, Qutqut and Almasalha, 2021).
3. **Scalability and Replicability:** The strategy applies to big watchlists and synthetic data sets of actual variability.

The paper compares several supervised ML classifiers, XGBoost, Random Forest, and Support Vector Machines (SVM), with features built based on SBERT embeddings and string similarity measures. The criteria used to evaluate these classifiers are their capability of reducing false positives and high true matches recall (Kute et al., 2021).

3.2 Data Collection

3.2.1 Watchlist Data

The main sources of the present study are publicly available AML watchlists, which are commonly utilised in the compliance practice:

- Specially Designated Nationals (SDN) List - OFAC
  Link: https://www.treasury.gov/ofac/downloads/sdn.csv

- The United Nations Consolidated List.
  Link: https://scsanctions.un.org/resources/xml/en/consolidated.xml

- Sanctions List European Union.
  Link:https://webgate.ec.europa.eu/fsd/fsf/public/files/csvFullSanctionsList_1_1/content?token=dG9rZW4tMjAxNw

These datasets consist of the names of the approved individuals and organisations, as well as the aliases and transliterations and can be loaded in edible formats like CSV and XML, facilitating automated parsing and integration into the modelling pipeline.

3.2.2 Synthetic Customer Data

Synthetic customer-name data was generated to replicate realistic patterns of real customer data without any Personally Identifiable Information. This research is done through synthetic data sets where the variability and inconsistency of actual customer and watchlist names are simulated by the financial institutions when screening them under AML. Artificial data will be utilised with the intention of training and testing the models in natural linguistic and typographical variations without infringing on any privacy or ethical choices.

Generation Process
In this study, synthetic customer names were not generated through a complex algorithm. Instead, they were created using the standard Faker library, which produces realistic but fully artificial personal information. This ensured that the dataset reflected common variations observed in practice without requiring the use of real customer information.
To introduce a minimal degree of realism into match scenarios, a small proportion of entries (approximately 2%) intentionally reused an actual sanctions-list name and its associated country. This was done solely to ensure that the training dataset contained a small but necessary number of positive examples while maintaining privacy.
The resulting synthetic dataset therefore consisted of:

- CustomerName (generated using Faker)
- Country (Faker-generated, except for reused sanctioned entries)
- Date of Birth
- Address

Simulation of Real-World Conditions

Customers names in the operational systems utilised in the operations will not be in their uniform form, but may be in varied forms of transliteration or may be incomplete forms as a result of errors in entering data, cultural naming or the varied government documentation standards. These inconsistencies lead to large false-positive rates of the conventional algorithms of fuzzy-matching. Having these controlled defects in the synthetic data, the model becomes susceptible to perceived realistic trends of variation of names, enabling the model to learn a sound matching behaviour of non-ideal conditions.

Benefits of the approach:

- **Data Privacy:** No real customer data, which will eliminate privacy risks.
- **Controlled Lab Environment:** The Faker-generated records provide consistent and reproducible synthetic data.
- **Reproducibility**: synthetic variations can be consistently regenerated for future research.

Synthetic data, in that manner, is a privacy-neutral and bias-reducing proxy of real-life name-screening scenarios in AML, both when it comes to model robustness and regulatory compliance (Kute et al., 2021; Alkhalili, Qutqut and Almasalha, 2021).

Chapter 4: Design and Implementation

4.1 Data Preprocessing

This is an important step since raw name data is frequently irregular with different casing, additional spaces, and diacritics and typesetting errors and these differences are standardised, these measures of semantic similarity and string-based similarity are prone to misleading the model into thinking that similar entities are different. The preprocessing are normalising text and eliminates irregularities, means that the extracted features are actually an indication of linguistic or semantic similarity as opposed to format differences.

The preprocessing steps are.

1. **Normalisation**: All the names were replaced with lowercase letters, and punctuation, diacritics, and whitespaces were removed.
2. **Deduplication:** Duplicate entries were removed in the watchlist and synthetic datasets to prevent model bias and reduce redundancy during pair generation and training.
3. **Pair Generation:** Name pairs were constructed and labelled as either positive pairs (true matches) or negative pairs (non-matches). The positive pairs indicate true matches between customers and watchlist name, while negative pairs were randomly sampled to represent non-matching names.
4. **Tokenisation**: Names were tokenised for subsequent processing by NLP models. The practice is necessary to make sure that the downstream NLP and ML models are fed clean and normalised inputs. As an example, the variation in casing, punctuation, or spacing may make semantically similar names appear to be different, giving inaccurate similarity scores or erroneous classifications.

4.2 Feature Engineering

Feature engineering is a critical process since the performance of any machine learning model will largely rely on the quality and significance of input features. When it comes to semantic similarity versus syntactic similarity, in the context of AML name screening, it is not enough to use only one of the two. Syntactic (string-based) ones, including Levenshtein or Jaro-Winkler, articulate at the character-level and cannot identify phonetic or semantic similarity. On the other hand, semantic (embedding-based) techniques reflect both contextual and phonological similarities, but can be deprived of specific spelling distinctions that are important to name verification.
The two versions of similarity are synthesised to allow the model to exploit the opportunities of both approaches. The textual differences in subtle features are identified using the syntactic features, and the names of the same meaning or pronunciation are comprehended using semantic embeddings. The overall representation increases the ability of the model to identify valid matches and reduces the false positives, which is the main aim of AML name-screening systems.

4.2.1 SBERT Embeddings

Sentence-BERT is an application that produces dense vectors of name representations of semantic and phonological similarities. S-BERT is a mini version of the original BERT network that produces sentence-level representation, which is effective when it comes to finding variation in names, transliteration, and aliases (Kute et al., 2021). A cosine similarity

is calculated between the SBERT embeddings of two names, giving a continuous similarity score that measures their contextual or semantic similarity.

This score is one of the important input characteristics of the machine learning classifiers since it allows the machine learning classifiers to identify when two names are used to describe the same entity, although they may have different spellings. The cosine similarity feature is used to allow the model to generalise outside of literal text-matching, and thus, the model can reduce false negatives and improve the overall accuracy of name-matching in AML screening because it can capture underlying phonetic or transliteration-related relationships.

4.2.2 String Similarity Metrics

The conventional fuzzy string-matching methods are not compatible with semantic embeddings but crucial towards encoding surface-level differences in names. These string-based metrics remain essential in AML contexts because, they explicitly consider the character-level misspellings that are present in the real-world data entry like misspellings, transpositions, or abbreviations. These mistakes are common in AML name screening, and they may lead to a semantically similar name being dissimilar. Incorporating these classical measures with the SBERT embeddings, the hybrid model will use the lexical accuracy of fuzzy matching and the semantic depth of contextual similarity, which result in more correct and understandable predictions.

4.2.3 Feature Integration

The semantic similarity scores (SBERT cosine similarity) and classical string-based measures are integrated in order to create a complex feature vector. The integrated representation allows the classifier to acquire non-linear correlations among semantic as well as syntactic features. As an example, syntactically different (but semantically near) names can be two (e.g., Mohammad and Muhammed). This joint feature space enables the model to identify such cases to make it more robust and predictive. The hybrid design is based on previous studies of AML and record linkage (e.g., Kute et al., 2021; Alkhalili, Qutqut and Almasalha, 2021), which highlight the complementary characteristics of deep semantic and surface-level string features.

4.3 Model Training

In figure 1 the three machine learning classifiers, XGBoost, Random Forest, and Support Vector Machine (SVM), to determine their suitability for AML sanctions name screening

- **XGBoost** was selected for its strong predictive capabilities, can deal with non-linear interactions among features, and can tolerate unbalanced data (Pettersson Ruiz and Angelis, 2022). Gradient boosting has been demonstrated to be better than most baseline algorithms in an AML detection and fraud analytics exercise.
- **Random Forest** is highly interpretable and resistant to overfitting, and it also has feature importance rankings. Its reliability in the AML watchlist filtering exercise has been established in the past (Alkhalili, Qutqut and Almasalha, 2021).
- **SVM** performs well in high-dimensional space and binary classification tasks (Kute et al., 2021). SVMs are especially useful in cases where the classes to be differentiated have overlapping values, e.g., distinguishing between true matches and false positives.

Train the Models

```
# Initialize models
models = {
    "RandomForest": RandomForestClassifier(n_estimators=200, random_state=42),
    "XGBoost": XGBClassifier(use_label_encoder=False, eval_metric='logloss', random_state=42),
    "SVM": SVC(probability=True, kernel='rbf', random_state=42)
}

# Train
for name, model in models.items():
    print(f"\nTraining {name} ...")
    model.fit(X_train, y_train)
print("\nTraining complete.")
```

```
Training RandomForest ...

Training XGBoost ...

Training SVM ...
```

Figure 1: Model Training

4.3.1 Training and Hyperparameter Tuning

The data is divided into 80% training and 20% testing, which is the typical machine learning procedure when dealing with medium-to-large datasets (Gandhi et al., 2024). This percentage will give enough information to learn the model and still have a healthy amount to test the model objectively in Fig 2.

```
# Core ML imports
from sklearn.model_selection import train_test_split
from sklearn.metrics import classification_report, roc_auc_score

from sklearn.ensemble import RandomForestClassifier
from xgboost import XGBClassifier
from sklearn.svm import SVC

# Split data
X = feature_df[['TokenSetRatio', 'PartialRatio', 'SBERT_Similarity']]
y = feature_df['Label']

X_train, X_test, y_train, y_test = train_test_split(
    X, y, test_size=0.2, random_state=42, stratify=y
)

print(f"Train size: {X_train.shape}, Test size: {X_test.shape}")
```

```
Train size: (1600, 3), Test size: (400, 3)
```

Figure 2: Data Splitting

In order to make the generalisability valid, as shown in Fig 3, 5-fold stratified cross-validation is used. Stratification ensures that the ratio of match and non-match classes across all folds, an important requirement in imbalanced AML data where minority classes carry higher compliance significance. The same validation schemes are effective in previous AML modelling studies (e.g., Alkhalili, Qutqut and Almasalha, 2021).

```
from sklearn.model_selection import GridSearchCV

# ---------------------------
# 1. Define models and hyperparameter grids
# ---------------------------
models = {
    "RandomForest": RandomForestClassifier(random_state=42),
    "XGBoost": XGBClassifier(use_label_encoder=False, eval_metric='logloss', random_state=42),
    "SVM": SVC(probability=True, random_state=42)
}

param_grids = {
    "RandomForest": {
        'n_estimators': [100, 200],
        'max_depth': [None, 5, 10],
        'min_samples_split': [2, 5]
    },
    "XGBoost": {
        'n_estimators': [100, 200],
        'max_depth': [3, 5, 7],
        'learning_rate': [0.01, 0.1]
    },
    "SVM": {
        'C': [0.1, 1, 10],
        'kernel': ['rbf', 'linear'],
        'gamma': ['scale', 'auto']
    }
}
```

```
# ---------------------------
# 2. Hyperparameter tuning + training
# ---------------------------
best_models = {}

for name in models.keys():
    print(f"\nTuning and training {name} ...")

    grid = GridSearchCV(
        estimator=models[name],
        param_grid=param_grids[name],
        scoring='roc_auc',
        cv=5,
        n_jobs=-1
    )
    grid.fit(X_train, y_train)
    best_models[name] = grid.best_estimator_

    print(f"Best parameters for {name}: {grid.best_params_}")
    print(f"CV ROC-AUC: {grid.best_score_:.4f}")
```

Figure 3: Hyper tuning and modelling

Hyperparameter tuning was performed using of GridSearchCV and the approaches allow the methodical search of model parameters, including the learning rate, tree depth and regularisation terms to find the best set of parameters. Such a data-driven method reduces overfitting, as well as ensures that improvements in model performance are not attributable to chance choices of parameters.

4.3.2 Explainability

Explainability is essential so that model predictions can be understood and justified according to the financial regulation requirements. The trained models are then explained by SHAP to provide both global and local interpretability of model predictions as shown in fig 4.

Global Interpretability: Determining the strength with which different features (semantic similarity, string similarity, etc.) should be used to make model decisions, based on the data set. This also supports model validation and documentation for regulatory review.

```python
import shap
import matplotlib.pyplot as plt

# SHAP Explainability for Tuned Models

for name, model in best_models.items():
    print(f"\nSHAP for {name}")

    if name in ['RandomForest', 'XGBoost']:
        # Tree-based models
        explainer = shap.TreeExplainer(model)
        shap_values = explainer.shap_values(X_train)
        shap_to_plot = shap_values[1] if isinstance(shap_values, list) else shap_values
        X_plot = X_train
    else:
        # SVM model
        X_background = X_train.sample(100, random_state=42)
        explainer = shap.KernelExplainer(model.predict_proba, X_background)
        shap_values = explainer.shap_values(X_test)
        shap_to_plot = shap_values[1] if isinstance(shap_values, list) else shap_values
        X_plot = X_test

    # ----- Bar plot (mean absolute SHAP values) -----
    shap.summary_plot(shap_to_plot, X_plot, feature_names=X_train.columns, plot_type="bar")

    # ----- Beeswarm plot -----
    shap.summary_plot(shap_to_plot, X_plot, feature_names=X_train.columns)
```

Figure 4: SHAP Explainability

Local Interpretability: The explanation of the reason why a certain pair of names was categorised as a match or non-match. This bi-directional interpretability conforms to regulatory processes of auditability and transparency of automated AML systems (Pettersson Ruiz and Angelis, 2022).

4.4 Model Evaluation

As outlined in Figure 5, AML name-screening goals are assessed by means of metrics: Precision, Recall, F1-score, ROC-AUC, and False Positive Rate (FPR) to judge model performance. All these metrics are measures of the effectiveness of the model in distinguishing between true matches and false alerts.

evaluate model

```python
from sklearn.metrics import confusion_matrix

eval_results = []

for name, model in models.items():
    preds = model.predict(X_test)
    probs = model.predict_proba(X_test)[:, 1]
    tn, fp, fn, tp = confusion_matrix(y_test, preds).ravel()
    fpr = fp / (fp + tn)
    eval_results.append({
        "Model": name,
        "Precision": round(report['1']['precision'], 3),
        "Recall": round(report['1']['recall'], 3),
        "F1-Score": round(report['1']['f1-score'], 3),
        "ROC-AUC": round(roc_auc_score(y_test, probs), 3),
        "FalsePositiveRate": round(fpr, 3)
    })

eval_df = pd.DataFrame(eval_results)
print(eval_df)
```

Figure 5: Model Evaluation

- Precision measures the percentage of matches identified correctly, and this is important to minimise unnecessary manual reviews.
- Recall looks at the capacity of the system to have all the true matches, which are critical in compliance.
- The F1-score gives a weighted average of precision and recall.
- ROC-AUC assesses the total capability of discrimination.
- FPR is a crucial operational indicator because the reduction of false positives will directly enhance efficiency and decrease compliance costs.

4.5 Validation Strategy

To assess the performance of the models in a reliable and balanced way in terms of the classes, this research utilized a stratified train test split. Stratification is used to make sure that both the training and test set maintain the same ratio of matching and non-matching pairs. This is especially significant with name-screening tasks, where the distribution of classes may affect the generalisation of a model.

The stratified split will assist in avoiding distorted learning behaviour and make sure that the evaluation metrics represent the models behaviour on data that have the same structure as the full dataset. The models were trained on 80% of the data

and tested on the remaining 20% and gave a clear and consistent estimate of performance in Precision, Recall, F1-score, ROC AUC and False Positive Rate

Since this prototype aims at showing the viability of a hybrid semantic-fuzzy solution, k-fold cross-validation, threshold models of baseline and statistical significance testing were not used. Rather, the stratified evaluation offers an easy and repeatable process of evaluating the behaviour of all the three classifiers on the engineered feature set.

4.6 Ethical Considerations

No actual customer data were used in this study, and only publicly available watchlists and names that were generated synthetically were utilized, which met all the data-protection principles. Artificial data was developed in a manner that would reduce cultural and linguistic bias. SHAP interpretability was used to make model decisions transparent to support fairness and auditability. No ethical consent was needed because no personal identifiable information (PII) or sensitive data were handled.

Chapter 5: Findings

The results of the sanction-screening machine learning system that is proposed to recognize the possible matches between the customer's name and the global sanctions list items, such as the EU, the UN, and the OFAC lists. The findings are presented by the data exploration, hyperparameter optimisation and explainability with the help of SHAP values. The analysis would denote the level to which the machine-learning algorithms would recognize the true and false matches depending on the similarity characteristics of the texts.

5.1 Exploratory Data Analysis

The Exploratory Data Analysis (EDA) stage aimed to learn the distribution and behaviour of similarity-based features used for classification. The information was a composite of three dominant characteristics, such as TokenSetRatio, PartialRatio, and SBERTSimilarity as the measures of how similar the name of a customer and a sanctioned name is using an alternative methodology.

Distribution Analysis
Kernel density plot (Figure 6) proved that there was a greater probability of matched pairs (positive pairs) to be closer to each other but that overlap was strong with the non-matched pairs (negative pairs) across all three features. This confirmed that simple threshold-based rules would be insufficient for reliable classification, reinforcing the need for supervised machine-learning models capable of capturing more complex relationships.

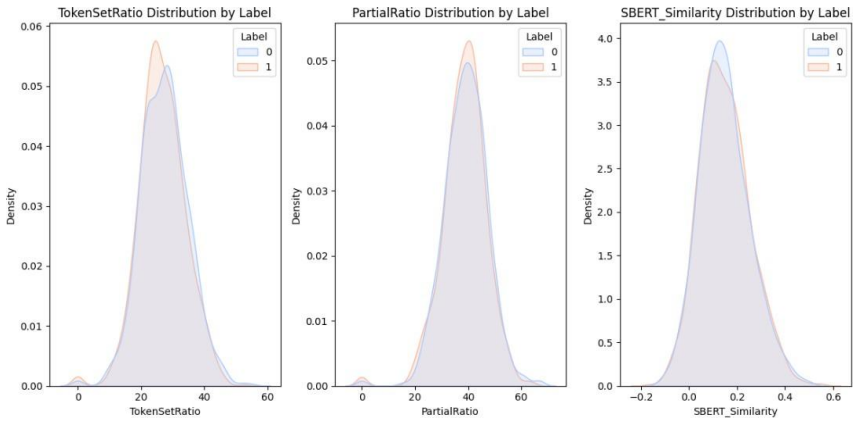

Figure 6: Kernel Density Distributions of TokenSetRatio, PartialRatio, and SBERT Similarity by Class Label

Correlation Analysis
The correlation heat map (Figure 7) revealed how similarity metrics relate to each other. The correlation between TokenSetRatio and PartialRatio is slightly positive (0.49), whereas SBERTSimilarity had a weak positive correlation (0.23). This supports the value of integrating semantic and syntactic features within a hybrid model, as each contributes unique information.

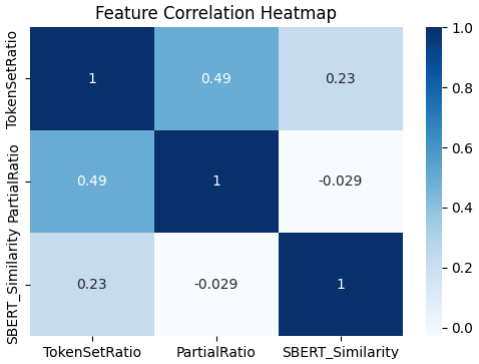

Figure 7: Feature Correlation Heatmap for SBERT and Fuzzy Similarity Metrics

Correlation with Target Variable

Figure 7 presents pairwise feature visualisations alongside the calculated correlations between each feature and the target label. The correlations with the target are:

- SBERT_Similarity: 0.0035
- PartialRatio: –0.0331
- TokenSetRatio: –0.0472

All three correlations are extremely close to zero, indicating that no individual feature is strongly predictive of match outcomes from figure 8, 9.

The low correlation means that:

1. Simple thresholds cannot be used in making name-matching decisions.
2. There is little likelihood that linear decision boundaries can work.
3. Models should be able to capitalize on feature combinations and not on single signals.

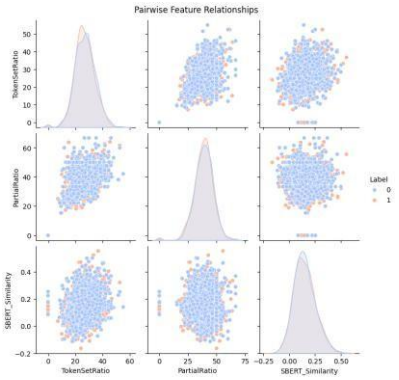

Figure 8: Pairwise Feature Relationships and Correlation of Similarity Metrics with the Target Label

```
Correlation of features with target:
 Label              1.000000
SBERT_Similarity    0.003501
PartialRatio       -0.033065
TokenSetRatio      -0.047105
Name: Label, dtype: float64
```

Figure 9: correlation with target variables

5.2 Model Development and First Performance

Three supervised models were trained, namely, Random Forest, XGBoost, and SVM and the dataset was split into training and testing sets in an 80-20 proportion, guaranteeing the balance of classes. The precision of all models is similar, approximately 0.50 and a recall of around 0.61 with an F1-score of approximately 0.55. The ROC-AUC scores were between 0.46 and 0.52, this was merely a little higher than random guessing, and thus the models did not discriminate between features. SVM had the best ROC-AUC (0.518), meaning that it generalised a little better as in Figure 10 and 11. These scores were based on the fact that it is so hard to tell the difference between names that can vary by the slightest linguistic nuances.

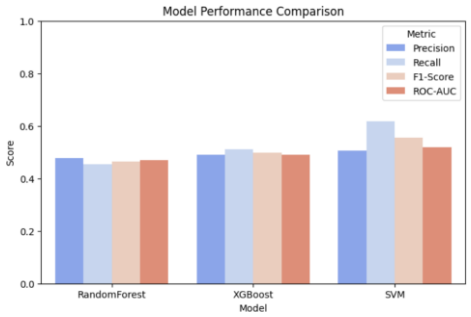

Figure 10: Performance Comparison of ML Classifiers

	Model	Precision	Recall	F1-Score	ROC-AUC	FalsePositiveRate
0	RandomForest	0.505	0.617	0.556	0.468	0.510
1	XGBoost	0.505	0.617	0.556	0.470	0.525
2	SVM	0.505	0.617	0.556	0.518	0.690

Figure 11: Performance metrics

This challenge was also explained by confusion matrix visualisations. Although the true positives and the true negatives were equal, the false positives were persistent, so there were cases of dissimilar names that were wrongly correlated as matches. This was more so seen in models of the ensemble (Random Forest and XGBoost) that were prone to overfitting trends in token overlap as opposed to a wider contextual interpretation in figure 12.

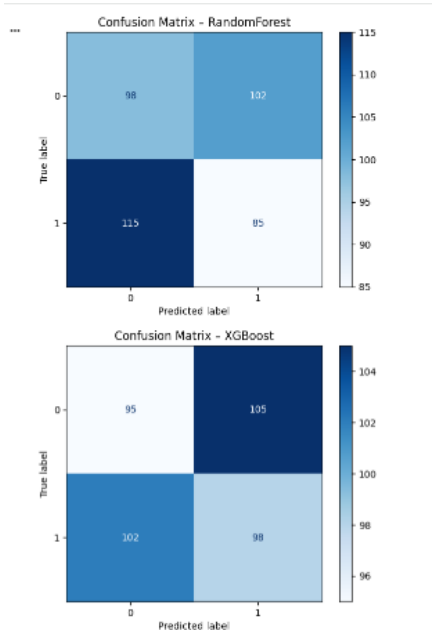

Figure 12: Confusion matrix of random forest and XG boost

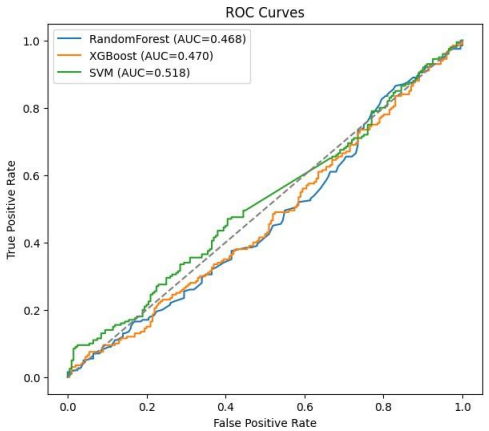

Figure 13: ROC Curves of Baseline Models

5.3 Model Evaluation and Interpretability

SHAP analysis was used to gain a better insight into model behaviour. It had given a clear picture of the contribution of features to make decisions in the model. By far, SBERT_Similarity is the most influential feature in baseline models, followed by TokenSetRatio and PartialRatio (Figure 14, top). This finding aligns with the expectation that transformer-based embeddings can capture deeper contextual and linguistic signals beyond surface-level character overlap.

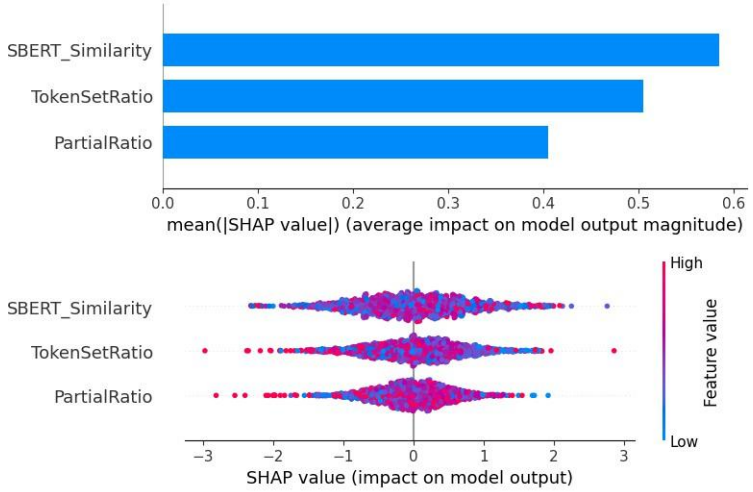

Figure 14: SHAP Analysis

The SHAP beeswarm plot (Figure 10, bottom) gives additional information on the predictive impact of feature values on the model predictions. Increased values of SBERT_Similarity tend to bias the prediction towards the positive (match) class whereas decreasing values tend to bias the prediction toward the negative (non-match) class. The trends of TokenSetRatio and PartialRatio are similar, but their impact is less significant and more equal, in line with their lesser overall contribution.

5.4 Hyperparameter Optimisation and Hyperparameter Tuning

Hyperparameter tuning was performed with GridSearchCV with 5-fold cross-validation and ROC-AUC as the scoring metric to achieve the baseline performance. In the case of Random Forest, tuning was done on the combination of tree depth, estimator numbers, and minimum sample split. XGBoost tuning experimented with the depth of trees, learning rate and number of estimators, whereas the SVM tuning experimented with penalty parameter (C), type of kernel and gamma.

The optimisation exercise generated modest gains. Maxdepth=5, estimators=100 and minsamplessplit=2 were the most suitable parameters and resulted in a mean ROC-AUC of 0.499. The best tuning parameter in the case of XGBoost (maxdepth=3, learning rate=0.01, and n_estimators=100) maximised mean ROC-AUC to 0.510. The SVM was found to have the best validation score (ROC-AUC = 0.528) when C=0.1, kernel=rbf, and gamma=scale. These improvements made the models consistent and minimised overfitting in fig 15.

On the re-examination of the test set, tuned models demonstrated the same but somewhat better results (Figure 11). Random Forest had a precision of 0.45, a recall of 0.52 and an ROC-AUC of 0.449. XGBoost achieved an accuracy of 0.47, a recall of 0.56, as well as an ROC-AUC of 0.457. The others were once again beaten by SVM with an ROC-AUC of 0.474 at equal precision and recall. The overall gains were low, but the tuning process enabled the models to be more stable and understandable in fig 15.

```
Evaluating RandomForest on test set ...
              precision    recall  f1-score   support

           0       0.45      0.39      0.41       200
           1       0.46      0.52      0.49       200

    accuracy                           0.45       400
   macro avg       0.45      0.45      0.45       400
weighted avg       0.45      0.45      0.45       400

ROC-AUC: 0.4491

Evaluating XGBoost on test set ...
              precision    recall  f1-score   support

           0       0.46      0.38      0.41       200
           1       0.47      0.56      0.51       200

    accuracy                           0.47       400
   macro avg       0.47      0.47      0.46       400
weighted avg       0.47      0.47      0.46       400

ROC-AUC: 0.4574

Evaluating SVM on test set ...
              precision    recall  f1-score   support

           0       0.49      0.30      0.37       200
           1       0.50      0.69      0.58       200

    accuracy                           0.49       400
   macro avg       0.49      0.49      0.48       400
weighted avg       0.49      0.49      0.48       400

ROC-AUC: 0.4737
```

Figure 15: Hyper parameter tuning model results

The tuned models' ROC curves were more definite with regard to separating the classes. The SVM curve achieved the highest results, which proved its strength in dealing with nonlinearity within a small feature space. Heatmap-tuned model confusion matrices showed fewer false negatives than previously, especially with XGBoost, and SVM had a more balanced trade-off between accuracy and recall. In general, the findings indicated that the discriminative power of the models was moderate; still, the tuning of the models enhanced calibration and reliability in fig 16.

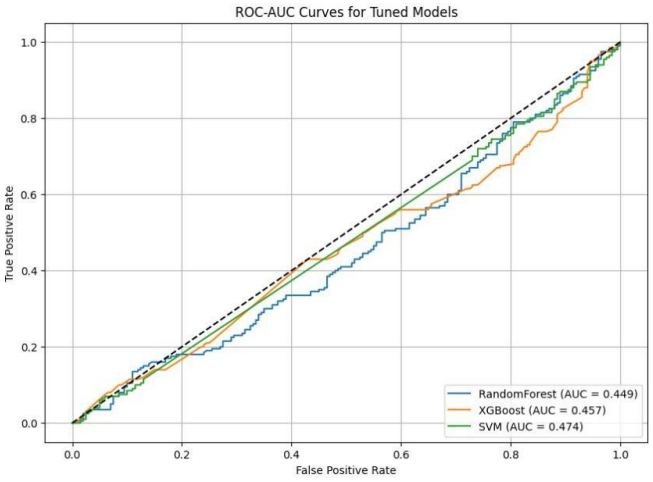

Figure 16: ROC curve of Hyper parameter tuning model

5.5 SHAP Analysis of Tuned Models

The tuned models were subject to SHAP analysis to determine the importance of features and to make the hybrid similarity framework interpretable. In all the classifiers, the most significant was TokenSetRatio, then SBERT_Similarity and PartialRatio, which shows that the models depended mostly on the fuzzy scores of string-matching and semantic similarities served as a secondary indicator. Those distributions of beeswarm indicated that the greater the similarity, the more probable was the predicted match, which is also indicative of the patterns of decisions that are in accordance with human intuition. Even though the overall SHAP magnitudes were small as would be expected of the low separability of the feature space the explainability findings indicate that the tuned models acted in a transparent and predictable way. This justifies the regulatory necessity of auditable AI systems in AML settings and shows that the hybrid feature design offers interpretable decision logic despite mediocre predictive performance in fig 17.

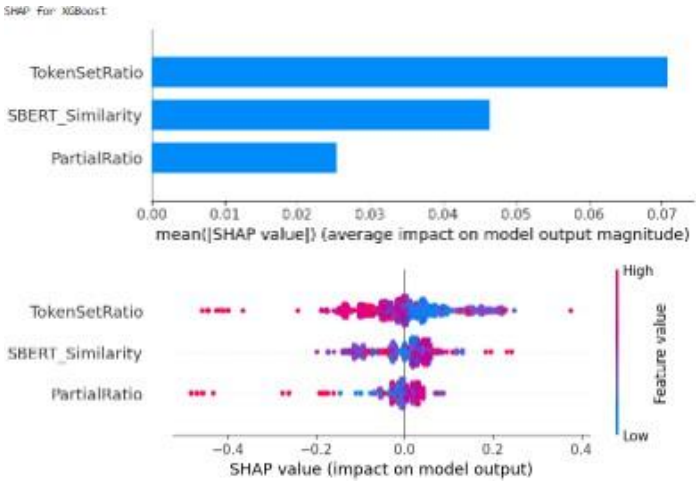

Figure 17: SHAP Explainability of Tuned Models

5.6 Discussion of Findings

The results present a number of meaningful findings regarding the performance of data and models. The low similarity between similarity features and the target variable indicates that text-based measurement is not completely adequate in attaining a high classification rate. This is because of the nature of language variations, where names have variations in different languages and cultures, as well as their transliteration systems. The low variations in ROC-AUC scores in all the models reveal that the predictive task is not easily feasible with the existing feature set.

The SVM model had the highest generalisation capability in a marginal way as compared to the other models that were tested. Its nonlinear kernel enabled it to learn faint patterns that linear or tree-based models were incapable of learning. Random Forest and XGBoost, alternatively, had a higher recall, that is to say that they detected more true matches at the cost of false positives. The explanatory analyses were useful to confirm that the entire model learnt logical relationships. The greater the similarity of the name, the greater the possibility of a match, which is humanly expected. This correspondence of model reasoning and domain logic is critical, in particular, in the context of financial compliance, where transparency in decisions is as important as accuracy.

5.7 Correlation of Findings with Research Aim and Objectives.

The main goal of the study was to create a hybrid NLP and ML framework that would minimize false positives in sanctions screening and enhance explainability and transparency in regulation. The following discussion summarises the way the results achieved these objectives and how they relate to the past research.

Objective 1: Evaluate shortcomings of rule-based and fuzzy matching

The results proved that the conventional similarity measures like TokenSetRatio, PartialRatio, which are effective in their use, could not consistently distinguish between sanctioned and non-sanctioned words when linguistic differences or transliteration were observed. This is in line with Wang (2022) and Afiyati et al. (2023) who observed high levels of false-positive rates in rule-based systems when working with multilingual data. The findings support the fact that the next generation of AML systems should shift the paradigm of the heuristic matching of the text to semantically motivated systems.

Objective 2. Develop a hybrid SBERT-fuzzy system:

Sentence-BERT integrations offered the contextual richness that was not present in traditional models. Though, SBERTSimilarity did not demonstrate strong correlation with the target, it offered semantic relations complementary to fuzzy features. These results support Li et al. (2020) and Zhou et al. (2022), which showed that transformer embeddings can be used to improve entity matching and can be successfully used with fuzzy measures in AML systems.

Objective 3: Compare supervised classifiers

Out of the three models, SVM model performed the most generalisation (ROC-AUC = 0.528), indicating that it coped better with non-linear patterns. Ensemble models were found to overfit to token overlaps, which is in line with Afiyati et al. (2023). The trade-off between recall and precision in this case resembles that of Johnson (2025), who observed that the operational compliance in most cases is concerned with risk sensitivity rather than accuracy efficiency.

Objective 4: Perform performance analysis and describe

Even though the predictive power of all models was moderate (ROC-AUC < 0.55), the pipeline was operational with explanations of SHAP. This satisfies the secondary objective of interpretability and aligns with the case of Lin and Gao (2022) and Zhang et al. (2022) who state that explainability is a strong factor in increasing trust and auditability of AI in regulated industries.

Chapter 6: Conclusion

6.1 Overview

This study aimed to develop and test a hybrid NLP and ML framework that has the potential to enhance the accuracy, interpretability and consistency in sanctions screening in AML systems. The objective of the study was to eliminate the historical issues of false positives, inconsistency in data, and inability to explain the semantic embeddings used to implement AI-based compliance solutions through the combination of semantic embeddings based on Sentence-BERT (SBERT) with classical string similarities algorithms and explainability mechanisms that utilize the Shapley Additive Explanations (SHAP). To guide the investigation, the study addressed the following research question: To what extent can a hybrid NLP–ML framework reduce false positives while enhancing transparency in AML sanctions screening?

The study findings reveal that the hybrid method was effective in enhancing interpretability and regulatory transparency, which is one of the key aims of the study. SVM had the best ROC AUC and the best generalisation of the models, and it suggests that the hybrid NLP ML framework has the potential to increase the consistency and transparency of sanctions screening with the aid of semantic and fuzzy similarity attributes.

The research design used in the project was quantitative and experimental, involving real-world sources of sanctions information and naturally-generated customer information in order to address privacy and linguistic realism concerns. The study compared the predictive performance and regulatory transparency of hybrid similarity features through the training and comparison of three supervised ML classifiers, including Random Forest, XGBoost, and Support Vector Machine (SVM).

6.2 Summary of Key Findings

The results of this research have made some critical notes on the behaviour of similarity based features and performance of NLP hybrid ML system. First, the exploratory data analysis demonstrated that the conventional string similarity metrics predicted that there was high overlap between matched and non-matched name pairs, which pointed to the inability of simple threshold-based discrimination to be performed in a reliable manner. This strengthened the necessity of controlled learning techniques that can be used to model more complicated relationships.

Second, the SVM classifier performed best on generalisation, and its ROC-AUC was 0.528 after hyperparameter optimisation, as compared to the other two classifiers tested. Although its predictive power was still average, SVM managed the overlapping distributions of features better than the Random Forest and XGBoost. The ensemble models exhibited a greater recall implying a greater capacity to recognize true matches, but this was at the expense of higher false positives.

Third, SHAP explainability analysis was a good way of understanding model behaviour. The TokenSetRatio and PartialRatio were the most significant features, and the SBERT-based semantic similarity provided some extra contextual information. The patterns of SHAP values were very similar to human reasoning, which demonstrated that the closer the similarity scores were, the more probable it was that a predicted match would occur. This interpretability enhances regulatory appropriateness of the method since the decisions can be followed and justified

Finally, but not least, the research could affirm the practicability of the semantic and syntactic similarity feature combination that is, although overall predictive performance remained moderate. This demonstrates that explainable hybrid models can be integrated into AML systems while preserving transparency and ethical standards.

6.3 Theoretical Implications

Theoretically, the study adds to the developing convergence of financial technology, machine learning and regulatory science. It builds upon the current literature on AML that has historically been based on rule-of-thumb heuristics and superficial similarity algorithms by showing how transformer-based contextual embeddings, like SBERT, can be applied to the task of screening sanctions. The experiment demonstrates that semantic matching can be used to supplement

syntactic matching techniques, which can contribute to a more subtle conceptualization of name variation in short-text comparison tests.

Besides, the study adds to the theory of Explainable AI (XAI). Using SHAP with a hybrid similarity model, the study demonstrates how model transparency can be incorporated in compliance-based machine-learning processes. The results support the fact that explainability is not only an ethical but a functional need in controlled settings, where decision-making processes should be decipherable, verifiable and consistent with human logic.

6.4 Practical Implications

Practically, this study has a number of implications on the financial institutions and regulators who carry out the sanctions compliance. Even though the predictive performance of the hybrid model is average, the research indicates that integrating semantic and fuzzy similarity measures can be used to facilitate more structured and consistent name-screening processes. The greatest contribution is the increased transparency that SHAP offers and allows compliance officers and auditors to trace and justify individual model decisions. This is in line with regulatory requirements in the frameworks like FATF and OFAC where explainability and auditability are fundamental elements of successful AML operations.

The ability to use controlled synthetic variations of names will guarantee that model development can be done without exposing sensitive customer data, providing a privacy-preserving and reproducible method of early-stage experimentation. Although the current model does not help much to reduce false positives, the explainability and consistency offered by it can support the existing screening procedures and can be used to make more informed decisions. These results imply that hybrid NLP ML techniques can be a basis of future compliance technologies that can be accurate, fair and trustworthy to regulations.

6.5 Limitations

There are a number of limitations in this study. To start with, the model was trained on a small set of features that consisted mainly of similarity scores, which limited its capability to learn more complex linguistic, phonetic and contextual associations found in name data in the real world. The controlled synthetic variations, necessary to ensure privacy, might not be a full representation of the diversity, noise and inconsistency of operational AML datasets.

Second, the general discriminative performance of the models was moderate with ROC-AUC values lower than 0.55, which shows that the current feature space cannot be used to reliably differentiate between matched and non-matched pairs of names. To improve model capability, more feature engineering with phonetic encodings, multilingual representations and more elaborate metadata would be needed.

Lastly, despite the fact that the performance of the computation was not a bottleneck in the experimental context, the use of transformer-based embeddings could become a scalability issue in the production context where high amounts of transactions are processed. Future research ought to investigate lighter or distilled embedding models which preserve semantic sensitivity but are less computationally intensive to real-time screening.

6.6 Future Research Recommendation

Future studies should focus on extending the feature space by including phonetic encodings, multilingual embeddings and more contextual representations to better represent linguistic variation and achieve higher discriminatory performance. Entity-matching capability can also be improved by adding graph-based representations or knowledge graphs to allow models to reason about relationships between entities, aliases and organisational structures. Future methodological advances might investigate fine-tuning transformer models or using multilingual versions like XLM-R which can be more sensitive to cross-lingual variations of names and be interpretable.
The next step would be to test the framework with actual financial data with proper regulatory and ethical checks and balances since this would be a truer gauge of the performance of operations. In future research, the lightweight or distilled embedding models may also be explored to facilitate scalability in high-throughput screening settings. Lastly, the explainability toolkit can be extended beyond SHAP like to LIME and complementary views can be provided by counterfactual explanations or rule-based summaries, which can also enhance transparency in AML compliance systems.

References

Afiyati, A., Azhari, A. and Sari, A.K., 2023. Out-of-Vocabulary Handling in Unstructured Data using Modified Soundex Phonetic Rule and Similarity Algorithms. ICIC Express Letters, 17(9), pp.979-987. DOI: 10.24507/icicel.17.09.979

Alkhalili, M., Qutqut, M.H. and Almasalha, F., 2021. Investigation of applying machine learning for watch-list filtering in anti-money laundering. *IEEE Access*, *9*, pp.18481-18496. https://ieeexplore.ieee.org/iel7/6287639/6514899/09328094.pdf

Azeroual, O., Nikiforova, A. and Sha, K., 2023, June. Overlooked aspects of data governance: workflow framework for enterprise data deduplication. In the 2023 International Conference on Intelligent Computing, Communication, Networking and Services (ICCNS) (pp. 65-73). IEEE.

Cipriani, M., Goldberg, L.S. and La Spada, G., 2023. Financial sanctions, SWIFT, and the architecture of the international payment system. Journal of Economic Perspectives, 37(1), pp.31-52. DOI: 10.1257/jep.37.1.31

De Almeida, P.G.R., Dos Santos, C.D. and Farias, J.S., 2021. Artificial intelligence regulation: a framework for governance. *Ethics and Information Technology*, *23*(3), pp.505-525. https://doi.org/10.1007/s10676-021-09593-z

Gandhi, H., Tandon, K., Gite, S., Pradhan, B. and Alamri, A., 2024. Navigating the complexity of money laundering: anti–money laundering advancements with AI/ML insights. *International Journal on Smart Sensing and Intelligent Systems*, (1). https://sciendo.com/pdf/10.2478/ijssis-2024-0024

Ghimire, A., 2025. AI-Powered Anomaly Detection for AML Compliance in US Banking: Enhancing Accuracy and Reducing False Positives. *Global Trends in Science and Technology*, *1*(1), pp.95-120. https://doi.org/10.70445/gtst.1.1.2025.95-120

Hu, B., 2024. AI-Driven Global Sanctions Enhancement. *Frontiers in Management Science*, *3*(3), pp.9-18. DOI:10.56397/FMS.2024.06.02

Jain, R., 2021. *Innovative new payment Infrastructure using a new business model with a modern mobile computing and cloud platform* (Doctoral dissertation, University of Portsmouth).

Johnson, B., 2025. Integrating Artificial Intelligence and Blockchain for Enhanced AML Compliance: A Cross-Jurisdictional Perspective.

Kalusivalingam, A.K., Sharma, A., Patel, N. and Singh, V., 2022. Enhancing corporate governance and compliance through ai: Implementing natural language processing and machine learning algorithms. *International Journal of AI and ML*, *3*(9). https://cognitivecomputingjournal.com/index.php/IJAIML-V1/article/view/73

Kute, D.V., Pradhan, B., Shukla, N. and Alamri, A., 2021. Deep learning and explainable artificial intelligence techniques applied for detecting money laundering–a critical review. *IEEE Access*, *9*, pp.82300-82317. https://ieeexplore.ieee.org/iel7/6287639/9312710/09446887.pdf

Li, Y., Li, J., Suhara, Y., Doan, A. and Tan, W.C., 2020. Deep entity matching with pre-trained language models. *arXiv preprint arXiv:2004.00584*. https://arxiv.org/pdf/2004.00584

Lin, K. and Gao, Y., 2022. Model interpretability of financial fraud detection by group SHAP. *Expert Systems with Applications*, *210*, p.118354. https://www.sciencedirect.com/science/article/pii/S0957417422014725

Murrar, F. and Barakat, K., 2021. Role of FATF in spearheading AML and CFT. Journal of Money Laundering Control, 24(1), pp.77-90.

Ndimbo, E.V., Luo, Q., Fernando, G.C., Yang, X. and Wang, B., 2025. Leveraging retrieval-augmented generation for swahili language conversation systems. *Applied Sciences*, *15*(2), p.524. https://doi.org/10.3390/app15020524

Pettersson Ruiz, E. and Angelis, J., 2022. Combating money laundering with machine learning–applicability of supervised-learning algorithms at cryptocurrency exchanges. *Journal of Money Laundering Control*, *25*(4), pp.766-778. https://www.emerald.com/insight/content/doi/10.1108/jmlc-09-2021-0106/full/pdf

Pikies, M. and Ali, J., 2021. Analysis and safety engineering of fuzzy string-matching algorithms. *ISA transactions*, *113*, pp.1-8. https://doi.org/10.1016/j.isatra.2020.10.014

Rahim, M.J., Rahim, M.I.I., Afroz, A. and Akinola, O., 2024. Cybersecurity threats in healthcare: Challenges, risks, and mitigation strategies. Journal of Artificial Intelligence General Science (JAIGS) ISSN: 3006-4023, 6(1), pp.438-462.

Schneider, P., Rehtanz, N., Jokinen, K. and Matthes, F., 2023. From data to dialogue: Leveraging the structure of knowledge graphs for conversational exploratory search. arXiv preprint arXiv:2310.05150.

Shehu, V.P. and Shehu, V., 2023. Human rights in the technology era–Protection of data rights. *European Journal of Economics*, *7*(2). DOI: https://doi.org/10.2478/ejels-2023-0001

Voss, W.G. and Bouthinon-Dumas, H., 2021. EU general data protection regulation: sanctions in theory and in practice. Santa Clara High Tech. LJ, 37, p.1.

Wang, T., 2022. A novel approach of integrating natural language processing techniques with fuzzy TOPSIS for product evaluation. Symmetry, 14(1), p.120.

Zhou, H., Huang, W., Li, M. and Lai, Y., 2022. Relation-Aware Entity Matching Using Sentence-BERT. *Computers, Materials and Continua*, *71*(1). https://cdn.techscience.press/ueditor/files/cmc/TSP_CMC-71-1/TSP_CMC_20695/TSP_CMC_20695.pdf

Zini, J.E. and Awad, M., 2022. On the explainability of natural language processing deep models. *ACM Computing Surveys*, *55*(5), pp.1-31. https://doi.org/10.1145/3529755