Şifrelenmiş Finansal Veriler ile Ensemble Öğrenme Yöntemleri Kullanılarak Bağımsız Denetim Görüşlerinin Çok Düzeyli Sınıflandırılması

Araştırma Makalesi/Research Article

Elif Nur KUCUR[1], Burak AKTÜRK[2], Ensar YILMAZ[2] , Tolga BÜYÜKTANIR[2]*, Kazım YILDIZ[1]

[1]Computer Engineering Department, Faculty of Technology, Marmara University, İstanbul, Türkiye
[2]Agra Research Lab, Agra Fintech Software Solutions, İstanbul, Türkiye

elifkucur@marun.edu.tr, burak.akturk@agrafintech.com, ensar.yilmaz@agrafintech.com, tolga.buyuktanir@agrafintech.com, kazim.yildiz@marmara.edu.tr

(Geliş/Received:29.05.2025; Kabul/Accepted:02.07.2025)
DOI: 10.17671/gazibtd.1708959

Özet— Bağımsız denetim raporları, şirketlerin finansal güvenilirliğini değerlendirmede kritik bir rol oynamaktadır. Denetçiler, görüşlerini finansal tabloların doğruluğu ve tutarlılığı ile bu tabloları oluşturan bileşenlere dayandırmaktadır. Bu çalışma, finansal tablolardan türetilen finansal oranlar ile Altman-Z, Springate ve Zmijewski gibi bilinen finansal risk skorlarını kullanarak denetim görüşlerini otomatik olarak sınıflandırmayı amaçlamaktadır. Sınıflandırma işlemi XGBoost ve Random Forest algoritmalarıyla gerçekleştirilmiştir. Veri gizliliği gereksinimleri göz önünde bulundurularak, modelleme süreci homomorfik şifrelemeyi destekleyen Concrete ML kütüphanesi kullanılarak yürütülmüş ve böylece finansal verilerin gizliliği korunmuştur. Nitelikli denetim görüşlerini daha ayrıntılı alt sınıflara ayırmak amacıyla hiyerarşik bir sınıflandırma yaklaşımı benimsenmiş, bu sayede yorumlanabilirlik artırılmıştır. Deneysel sonuçlar, önerilen modelin hem doğruluk hem de F1 skoru açısından güçlü bir performans sergilediğini göstermektedir. Geliştirilen sistemin, resmi denetim süreci öncesinde denetçilere ve diğer paydaşlara öngörüye dayalı, sistematik ve gizliliği koruyan bir karar destek mekanizması sunması beklenmektedir.

Anahtar Kelimeler— mahremiyet koruyan makine öğrenmesi, bağımsız denetim görüşü sınıflandırma, hiyerarşik sınıflandırma, finansal oranlar, topluluk öğrenmesi

Multi-Level Classification of Audit Opinions Using Ensemble Learning Methods with Encrypted Financial Data

Abstract— Independent audit reports play a crucial role in assessing the financial reliability of companies. Auditors base their opinions on the accuracy and consistency of financial statements and their underlying components. This study aims to automatically predict audit opinions by leveraging financial ratios derived from financial statements, as well as well-known financial risk scores such as Altman-Z, Springate, and Zmijewski. Classification was performed using XGBoost and Random Forest algorithms. Considering data privacy requirements, the modeling process was implemented using the Concrete ML library, which supports homomorphic encryption, thereby preserving the confidentiality of financial data. A hierarchical classification approach was adopted further to subdivide unqualified audit opinions into more detailed sub-classes, enhancing interpretability. Experimental results show that the proposed model achieves strong performance in terms of both accuracy and F1 score. The developed system is expected to serve as a predictive, systematic, and privacy-aware decision support tool for auditors and other stakeholders prior to the formal audit process.

Keywords— privacy-preserving machine learning, audit opinion classification, hierarchical classification, financial ratios, ensemble learning

1. INTRODUCTION

Independent audit reports are critical documents that reflect third-party evaluations of organizations' financial conditions. These reports typically include four main types of opinions: unqualified, qualified, adverse, and disclaimer of opinion. Audit opinions help evaluate the reliability of financial statements while also serving as an important decision-making tool for stakeholders such as investors, lenders, and regulatory bodies [1-4]. However, the fact that these processes involve subjective decisions, combined with factors such as time pressure, client relationships, and commercial competition, often raises the question of whether audit opinions carry absolute accuracy [5-7]. Especially, evidence in the literature highlights that the "disclaimer of opinion" is strategically used to conceal hidden negative aspects [8-10].

In this context, recent research has focused on the use of artificial intelligence and machine learning-based systems for the prediction and classification of independent audit opinions. Especially the studies conducted after 2023 have shown that methods such as XGBoost [11], LightGBM [12], Random Forest [13], Support Vector Machines [14-15], and deep learning [15-17] offer exceptional accuracy levels in predicting independent audit opinions. However, in solving this problem, not only strong classifiers but also data preprocessing, feature selection, and class imbalance mitigation strategies are of critical importance [14, 18-20].

Data imbalance is one of the fundamental issues encountered in the classification of audit opinions. Classes such as adverse and disclaimer of opinion typically constitute a very small percentage of the total dataset. This imbalance causes models to predominantly lean towards the unqualified opinion class and weakens their ability to differentiate between classes [21-22]. To address this issue, data augmentation algorithms such as SMOTE (Synthetic Minority Oversampling Technique) are used to increase the representation of minority classes and improve the model's generalization capacity [23-25]. Additionally, some studies have expanded these data augmentation processes by incorporating LLM (Large Language Models) supported methods, producing more realistic synthetic examples [26-28].

Additionally, recent research has brought the usability of hierarchical classification approaches in the context of independent auditing to the forefront. Especially, studies that categorize unqualified opinions into subcategories such as "Unqualified ++" and "Unqualified -" to analyze the risk level in more detail have emerged in the literature [29-30]. In this way, it becomes possible to express the degree of risk carried by the organization based on its financial structure, beyond the individual opinion classes. Such sub-classification models not only enhance classification accuracy but also improve the interpretability of audit decisions [31].

In this study, which aims to classify auditors' opinions based on both high-level and detailed sub-classes, the risk scores generated by normalizing different financial scores constitute the main input of the classification process [6,14]. In this framework, both hierarchical structure and SMOTE-supported imbalanced class modeling techniques have been combined. In practice, the XGBoost and Random Forest algorithms were preferred, and the classification performance was thoroughly evaluated using accuracy, F1 score, precision, and recall.

Our contribution aims to support independent auditors and stakeholders in making more informed auditing decisions before the actual audit process. By enabling faster, more systematic, and objective predictions of independent audit opinions, this study facilitates the early identification of potential outcomes. Additionally, it enhances the transparency and explainability of audit decisions, providing a scientific foundation for strengthening trust in the auditing process.

1.1. Research Contribution

This study offers the following contributions to the existing approaches in the literature regarding the prediction of independent audit opinions:

- A unique risk score was created by combining the Altman, Springate, and Zmijewski scores calculated using financial ratios. This score has been evaluated as an effective attribute in the sub-class definition process and has significantly increased the prediction success.
- Audit opinions have been modeled not only within four main categories but also divided into more detailed subcategories. Thus, intra-class heterogeneity has been reduced, and interpretability and prediction accuracy have been increased.
- The SMOTE method has been applied to address the issue of minority classes not being learned in imbalanced data structures. In this way, by ensuring balance between the classes, the model's sensitivity, especially for the minority classes, has been improved. Based on prior experimental studies conducted during earlier phases of our research and implementation, we observed that applying SMOTE consistently led to an improvement in model accuracy and overall performance, particularly in handling class imbalance [32].
- It has been demonstrated that the XGBoost algorithm is a strong option in terms of both classification performance and model stability. Although the training time is long, it has provided superior accuracy, especially in complex sub-classifications.

1.2. Paper Organization

This study consists of four main sections. The first section explains the main problem of the research, the importance of the topic, and the source of motivation. In line with the increasing machine learning-based studies in the post-2023 period, the current literature on the prediction of independent audit opinions is being examined in detail. Material and methods section includes the dataset used in the analysis, the financial indicators obtained from this dataset, and the risk scores generated. At the same time, the developed multi-level label structure is explained in this section. Then, the research methodology is defined, including data preprocessing steps, the imbalance correction process with SMOTE, and the application methods of classification algorithms. In the third section, experimental results are shared, and model performances are evaluated and interpreted in detail using accuracy, recall, precision, and F1 score. In the final section, the overall results of the study are summarized, the contribution to the literature is emphasized, and suggestions for future research are presented.

2. MATERIAL AND METHODS

2.1. Dataset Description

In this study, a publicly available dataset "Audit opinions of Turkish Public Companies" has been employed, which was taken from Kaggle [32]. The dataset compiles detailed information in terms of finance and corresponding audit opinion data for publicly listed companies on Borsa Istanbul (BIST). Dataset constructed based on publicly released quarterly reports and audit statements filed through the Public Disclosure Platform (KAP- Kamuyu Aydınlatma Platformu), according to accuracy and regulatory standards.

The database contains 2008-2024 years and firms, with 53 columns that track a wide range of financial measurements. These include financial statement items such as total equity, total assets, and total liabilities, and the breakdown of current and non-current assets and liabilities. Also, apart from basic financial values, the database includes ratio-based measures used in common applications in financial analysis. They are liquidity ratios (such as current ratio, quick ratio, and cash ratio), profitability ratios (such as net profit margin, operating margin, return on equity, and return on assets), and leverage ratios such as the debt-to-equity ratio and financial leverage. Furthermore, financial efficiency is represented through indicators like asset turnover, receivables turnover, and inventory turnover. One interesting aspect of the dataset is that it contains some of the bankruptcy prediction scores (such as Altman Z-Score, Springate Score, and Zmijewski Score), offering additional information related to the financial distress condition of each company. The target variable is opinion type on audit, which is categorized into classes as *unqualified*, *qualified*, *adverse*, and *disclaimer of* opinions. The auditor's opinions occurred with frequencies of 9610, 1054, 81, and 6, respectively.

2.2. Data Pre-processing

To ensure that the quality and consistency of the dataset are preserved prior to model training, several pre-processing data operations have been conducted.

2.2.1 Handling Missing Values and Feature Selection

Initially, all rows containing missing values were deleted from the dataset. Subsequently, a subset of features was dropped based on domain relevance and redundancy. Columns such as *company name*, *company code*, *period*, *year*, and various total aggregates (*e.g., total assets*, *total liabilities*, *total equity*) were excluded. These features were either identifiers, could introduce multicollinearity due to their derivation from other included features, or, particularly in the case of aggregate financial values, were more susceptible to inflationary distortions over time, potentially reducing the comparability across years.

2.2.2 Feature Scaling

After feature selection, all numerical features in the dataset were normalized according to the Min-Max Scaling method. It scales every numerical variable to a common scale between the range [0,1] and preserves the ratios among data points.

2.2.3 Label Encoding

The target variable, referring to audit opinion type, was initially categorical. To prepare it for use in classification models, it was converted into a numeric format through label encoding. The method assigns an integer label to each unique class such that the model can accurately interpret the categorical targets.

2.2.4 Data Splitting

Finally, the dataset was split into train and test sets according to the 80-20 ratio. The parameter random_state was passed to ensure reproducibility of the results.

2.2.5 SMOTE Oversampling

One of the key challenges in the dataset is the class imbalance problem, as the distribution of audit opinion

types is heavily skewed toward unqualified opinions. To address this issue and improve the classifier's ability to learn patterns associated with minority classes, SMOTE was applied to the training data. It operates by generating synthetic samples for the minority classes through interpolation between present instances and their closest neighbors, rather than by simple duplication. This method helps balance the class distribution without increasing the risk of overfitting[24]. By applying SMOTE, the dataset used for training becomes more representative, allowing classification models to better learn decision boundaries across all classes. In this study, SMOTE was applied before model training for both Random Forest and XGBoost, as well as for the privacy-preserving versions implemented using Concrete ML.

2.3. Concrete ML

Concrete ML is an open-source Python library that enables machine learning models to operate on encrypted data using fully homomorphic encryption. This technology allows model training and prediction even in environments where data remains encrypted, thereby preserving data privacy and enabling secure analysis. Concrete ML is designed to be compatible with popular machine learning libraries such as scikit-learn and PyTorch, allowing users to work with encrypted data without changing their existing workflows. It is particularly suitable for sectors with high security requirements, such as healthcare, finance, and defense, facilitating the development of machine learning applications without compromising data confidentiality [33].

2.3.1 Random Forest

Random Forest is a robust classification and regression algorithm developed within the scope of ensemble learning, which works by combining multiple decision trees. Each decision tree is trained with a randomly sampled subset from the original dataset. Thus, by training different trees, the model's generalization ability is enhanced and the risk of overfitting is reduced [14,34]. The basic prediction mechanism of the Random Forest algorithm is based on determining the final result according to the majority vote of each tree's decision. In the equation, y is calculated.

$$\hat{y} = mode(h1(x), h2(x), \ldots, hT(x)) \quad (1)$$

Here, h1(x) represents the prediction of the t-th decision tree, and T represents the total number of trees [42]. Especially in data structures where financial ratios are numerous and highly correlated with each other, Random Forest is quite advantageous. Each tree is trained with randomly selected subsets of features during training, which reduces the impact of inter-variable dependencies and allows the model to learn more diversely [35].

In this study, Random Forest was applied separately for both the main classes and the sub-opinion classes. Both the four main types of opinions and the sub-classes detailing positive opinions have been separately structured. To address the negative impact of data imbalance, especially on minority classes, synthetic examples were generated using the SMOTE method, and Random Forest training was conducted with these examples. The imbalance data issue was reduced with SMOTE, and it exhibited reasonable accuracy and F1 performance.

2.3.2 XGBoost

XGBoost is an optimized and highly efficient version of the gradient boosting method. This algorithm builds new trees by trying to minimize the errors of the previous model in each iteration. The main goal is to sequentially learn in a way that reduces the total loss function [11]. The general objective function of XGBoost can be defined as equation 2:

$$L(\phi) = \sum_{i=1}^{n} \quad l\left(y_i, \hat{y}_i^{(t)}\right) + \sum_{k=1}^{t} \quad \Omega(fk) \quad (2)$$

Here, the loss function is the regularization term that penalizes model complexity as Equation 3.

$$\Omega(\mathrm{f}) = \gamma \mathrm{T} + \frac{1}{2}\lambda \sum_{j=1}^{t} \quad w_j^2 \quad (3)$$

XGBoost prevents overfitting while maintaining the overall accuracy of the model and providing fast training and inference times [36-37].

XGBoost has been used in this study for multi-class classification, binary classification, and hierarchical structure.

In all these scenarios, data was balanced using the SMOTE algorithm, and then classification was performed with XGBoost. This structure has reduced the imbalance between classes, increasing both accuracy and the F1 scores for minority classes [34].

2.4. Proposed Method

The methodological framework established in this research focuses on enhancing precision, clarity, and neutrality in classifying audit opinions through structured financial information and calculated risk profiles. The overall process is presented in Figure 1 and starts from the integration of financial statements through their computation of financial ratios and finally concludes in the opinion prediction phase aimed at producing the expected audit opinion.

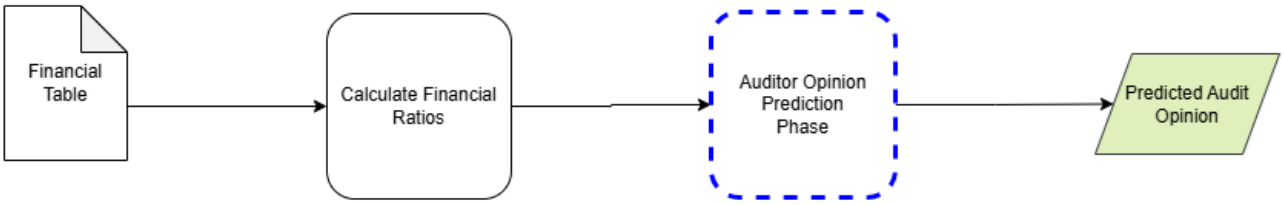

Figure 1. End-to-end auditor's opinion prediction processing pipeline from financial data

The two classification methods used at this opinion forecasting phase differ and consist of a hierarchical classification system, Phase A and Phase B, as presented in Figures 2 and 3, respectively.

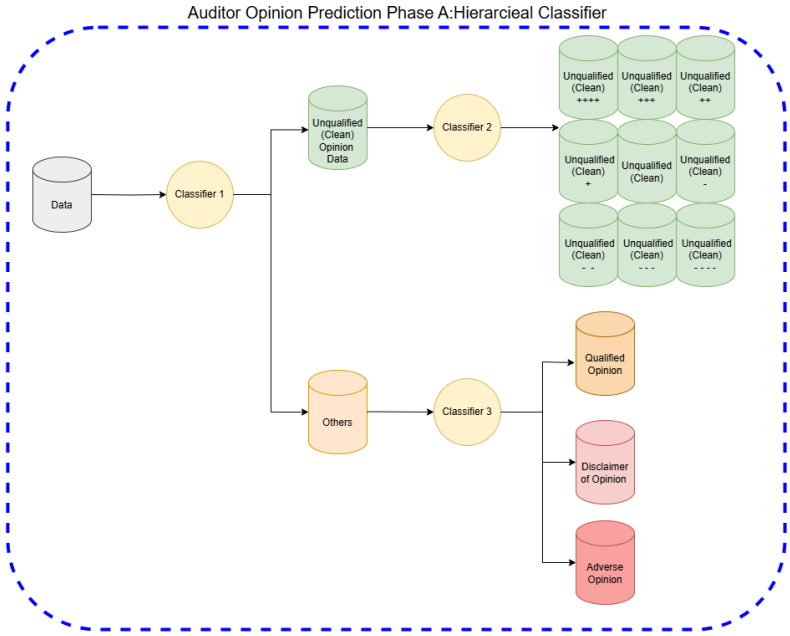

Figure 2. Hierarchical classification model for detailed auditor opinion prediction.

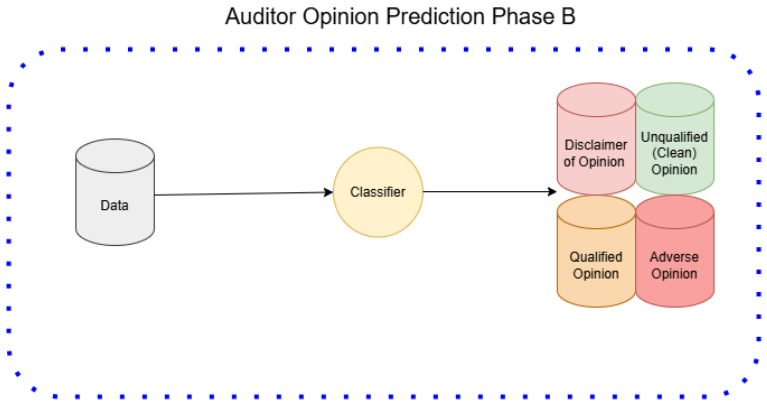

Figure 3. Flat classification model for direct auditor's opinion prediction.

In Phase A, a hierarchical classification framework is implemented to effectively capture the nuanced structure of audit opinions. Initially, the input data, comprising calculated risk scores and financial ratios, is evaluated using a sequence of three classifiers. The first classifier separates the data into two broad categories: instances likely to receive an Unqualified audit opinion and all other cases. For records classified as Unqualified, a second classifier assigns more granular opinion labels such as "Unqualified ++++," "Unqualified +++," "Unqualified ++," and so on. These subcategories are determined using weighted assignments and sigmoid-based scoring functions [38]. Instances not categorized as Unqualified are passed to a third classifier, which distinguishes among Qualified, Disclaimer of Opinion, and Adverse Opinion classes. To address class imbalance in this step, synthetic samples generated using the SMOTE method are employed, enhancing the model's ability to learn from under-represented classes.

Differing from Phase A, Phase B, illustrated in Figure 3, adopts a simpler, flat classification approach, where a single classifier directly predicts one of the four primary audit opinion types: Unqualified, Qualified, Adverse, or Disclaimer of Opinion. While this method is computationally more efficient and easier to implement, it lacks the multi-level interpretability and subclass refinement provided by the hierarchical structure in Phase A. Moreover, the flat model may be less robust in handling class imbalance and in accurately identifying less frequent opinion types, such as Adverse and Disclaimer of Opinion. This study aims to evaluate the effectiveness of hierarchical modeling in predicting audit opinions by comparing the classification accuracy and interpretability of the hierarchical approach (Phase A) with those of the flat approach (Phase B), particularly in addressing the challenges posed by imbalanced data and the complexity of financial reporting.

3. RESULTS AND DISCUSSION

3.1. Experimental Design

To evaluate the effectiveness of the proposed two-phase classification framework for predicting independent audit opinions, a comprehensive experimental setup was established. Both Phase A (hierarchical classification) and Phase B (flat classification) were implemented using four distinct machine learning models: Random Forest, XGBoost, Concrete Random Forest, and Concrete XGBoost. The Concrete-ML variants were utilized to enable privacy-preserving computations, supporting both training and inference on encrypted data.

Each model was applied under both Phase A and Phase B configurations to assess not only classification accuracy but also model interpretability and fairness, with particular emphasis on managing class imbalance.

In Phase A, the classification process involved a sequential chain of three classifiers. Classifiers 1 and 2 addressed relatively balanced classification tasks. However, Classifier 3 encountered significant class imbalance due to the rarity of certain opinion types. To mitigate this issue, the SMOTE algorithm was applied prior to training Classifier 3. This approach increased the representation of minority classes by generating synthetic samples, thereby enhancing the model's ability to learn more accurate decision boundaries.

3.2. Performance Metrics

In the evaluation of machine learning models, metrics perform a critical role in revealing the model's class discrimination capability, its success on minority classes, and its overall reliability [31,39-40]. In this study, the models developed for the classification of audit opinions have been analyzed within the framework of four basic metrics: Accuracy, Precision, Recall, and F1 score. Additionally, the meaning of the metrics, their mathematical definitions, and the reasons for their preference in this study have been explained in detail.

Accuracy measures the prediction accuracy of the model across all classes. TP (True Positive), TN (True Negative), FP (False Positive), and FN (False Negative) are calculated according to the classification results as shown in equation 4:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (4)$$

Although this metric indicates the overall success level, high accuracy can be misleading in datasets with imbalanced class distribution [37]. Therefore, other metrics have also been used in the performance evaluation of our study.

Precision measures how many of the examples classified as positive by the model are actually positive. In Equation 5, the calculation of the precision value is shown:

$$Precision = \frac{TP}{TP+FP} \quad (5)$$

This metric is particularly useful for measuring the quality of predictions made for minority classes, such as avoiding giving an opinion [37,41].

Recall shows how many of the truly positive examples are correctly classified. It means minimizing the FN value in equation 6:

$$Recall = \frac{TP}{TP+FN} \quad (6)$$

The avoidance of missing minority classes is particularly emphasized in areas where reliability is critical, such as auditing [37,41].

F1-Score is the harmonic mean of Precision and Recall and indicates the balanced performance of the model. It is calculated as in Equation 7:

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision+Recall} \quad (7)$$

This metric is widely used to express overall success, especially in imbalanced datasets [41].

3.3. Experimental Results

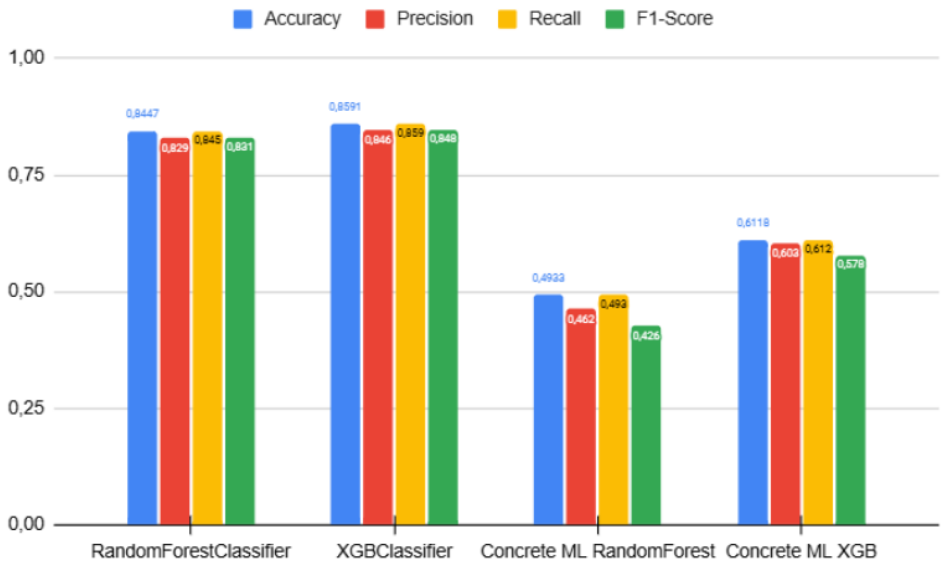

Figure 4. Classification results using all labels

The experiments were conducted using the complete set of audit opinion classes. As shown in Figure 4, which presents a comparative analysis of model performance, ensemble methods, namely RandomForestClassifier and XGBClassifier, consistently outperformed simpler baseline models across all major evaluation metrics, including accuracy, precision, recall, and F1-score.

Among the ensemble models, XGBClassifier achieved the highest accuracy at 0.859, followed closely by RandomForestClassifier with an accuracy of 0.845. Both models also demonstrated high precision and recall values, reflecting their effectiveness in accurately identifying instances across all audit opinion categories.

In contrast, the baseline models showed considerably lower performance, with accuracy scores of approximately 0.490 for RandomForest and 0.610 for XGBoost. These results indicate that simpler models struggle to capture the complexity and diversity of the dataset, leading to a decline in classification accuracy.

To further assess model behavior, confusion matrices for each configuration are presented in Figure 5. These matrices offer a detailed view of each model's ability to differentiate among the various audit opinion classes, highlighting both strengths and areas of misclassification.

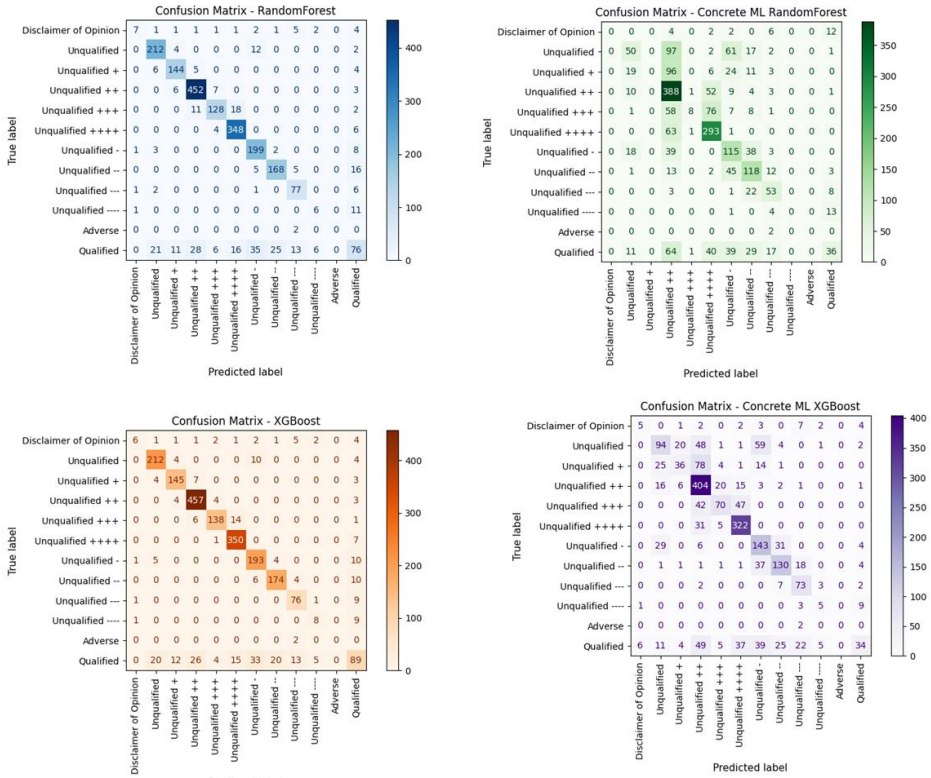

Figure 5. Confusion matrices obtained using all classes

3.3.1 Classifier-1 Results: *Unqualified* vs. *Others (Phase A)*

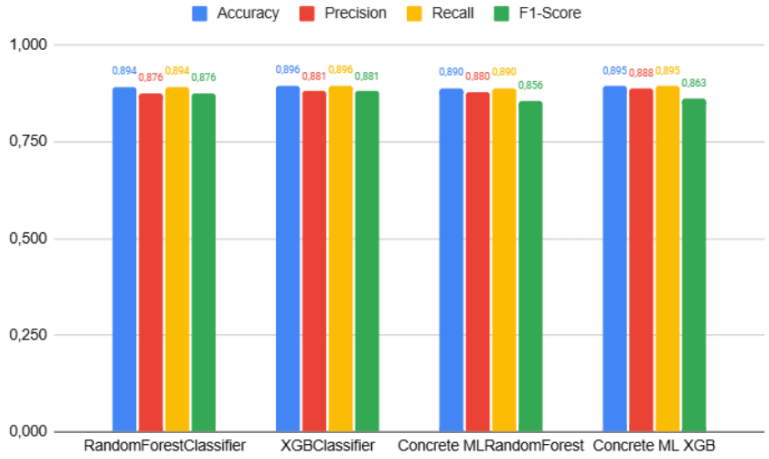

Figure 6. Classifier-1 results

At the first level of the hierarchical classification system (Classifier-1), the objective was to categorize audit opinions into two broad groups: Unqualified and Others. As shown in Figure 6, the XGBoostClassifier achieved the highest performance, with an accuracy of 0.896 and an F1-score of 0.881, indicating strong generalization capability. The RandomForestClassifier followed closely, with an accuracy of 0.894 and an F1-score of 0.876, demonstrating that both ensemble methods are well-suited to this binary classification task.

The encrypted models implemented using Concrete ML also produced encouraging results. Specifically, Concrete-ML XGBoost achieved an accuracy of 0.895 and an F1-score of 0.863, while Concrete-ML RandomForest recorded an accuracy of 0.890 and an F1-score of 0.856. These results suggest that the privacy-preserving models retain a high level of discriminative power, with only marginal decreases in performance compared to their unencrypted counterparts.

To further analyze model behavior at this stage, confusion matrices for each configuration are provided in Figure 7. These matrices offer a detailed view of how effectively each model distinguishes between Unqualified and Others categories during the initial step of the hierarchical classification process.

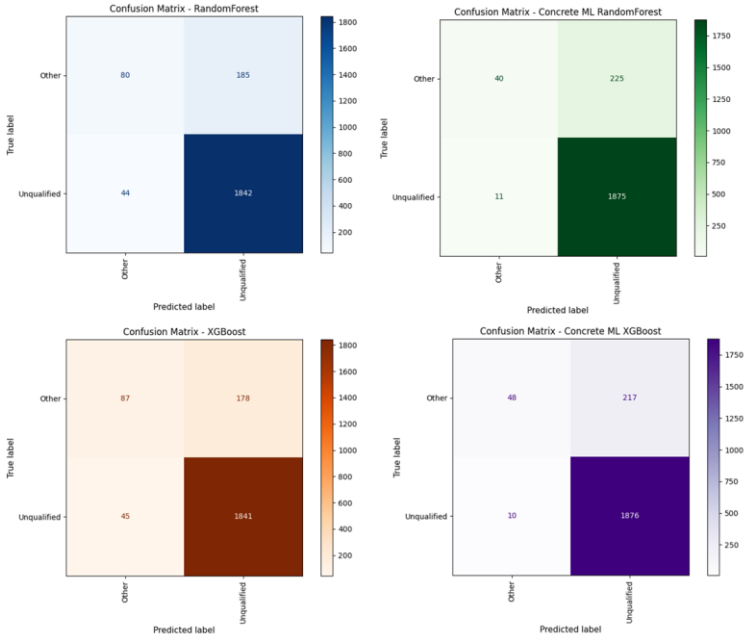

Figure 7. Confusion matrixes obtained from Classifier-1

3.3.2 Classifier-2 Results: Subclasses of *Unqualified Opinion (Phase A)*

Classifier-2 results (Figure-8) show that traditional models such as XGBoostClassifier and RandomForestClassifier achieve high consistency across all evaluation metrics. Both models deliver strong and stable predictions, indicating their effectiveness in this stage. In contrast, while the Concrete ML versions perform relatively lower, they still yield acceptable results considering the encrypted environment. Overall, this stage highlights clear model differentiation in terms of precision and robustness.

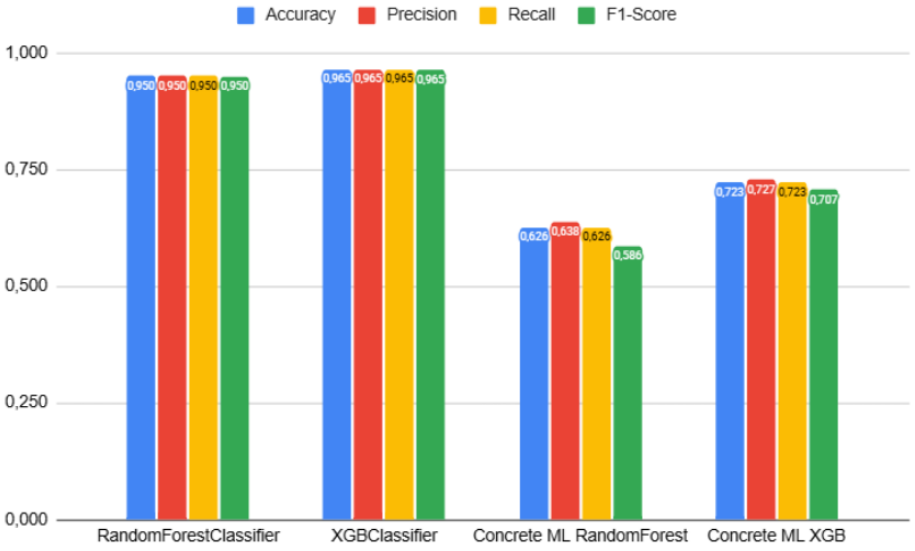

Figure 8. Classifier-2 results

For better understanding the performance of the models used by Classifier-2, the respective confusion matrices for each configuration are provided in Figure-9. These matrices describe the classification patterns demonstrated by the models as compared to the distribution of subcategories of the *Unqualified* class at the next level of the hierarchical classification system.

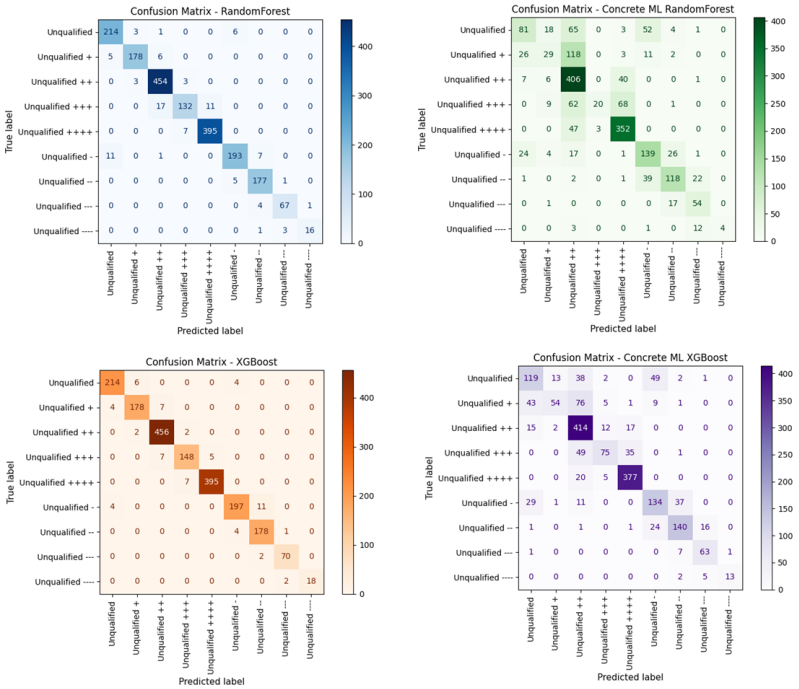

Figure 9. Confusion matrixes obtained from Classifier-2

3.3.3 Classifier-3 Results: Non-Unqualified Classes (Phase A)

The final level of the hierarchical system discerns between the three different types of non-qualified opinions as identified by Classifier-3. The results (Figure-10) show that both the RandomForestClassifier and the XGBoostClassifier perform well, as attested by very similar and stable metric values. In addition, the Concrete ML models achieve notable results, with the Concrete ML XGBoost having performance metrics very similar to those of its unencrypted counterpart. This phase demonstrates the models' ability to successfully deal with multi-class differentiation, even in an encrypted environment.

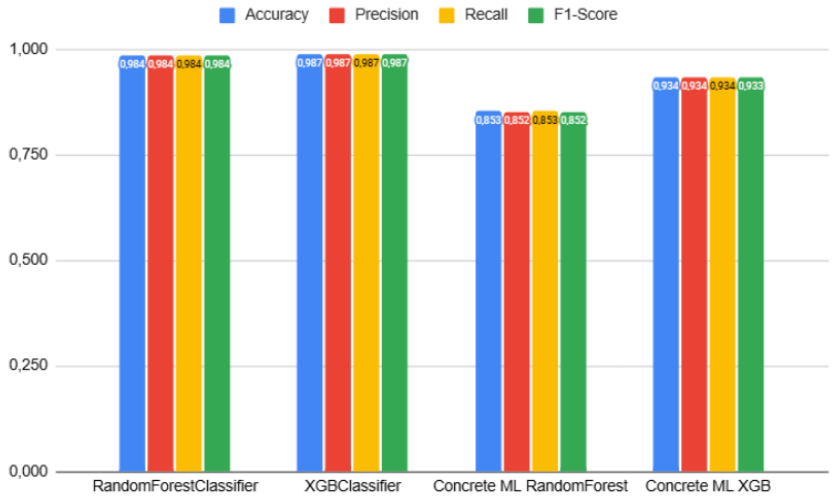

Figure 10. Classifier-3 results

As a means of clarifying the classifying performance of Classifier-3, confusion matrices for all of the models are shown in Figure-11. The matrices indicate the performance of the models for discrimination between the non-unqualified opinion categories at the last level of the hierarchical method. The graphical presentations affirm the extensive performance measures and attest the models' skill in dealing with multi-class distinctions through a high level of accuracy.

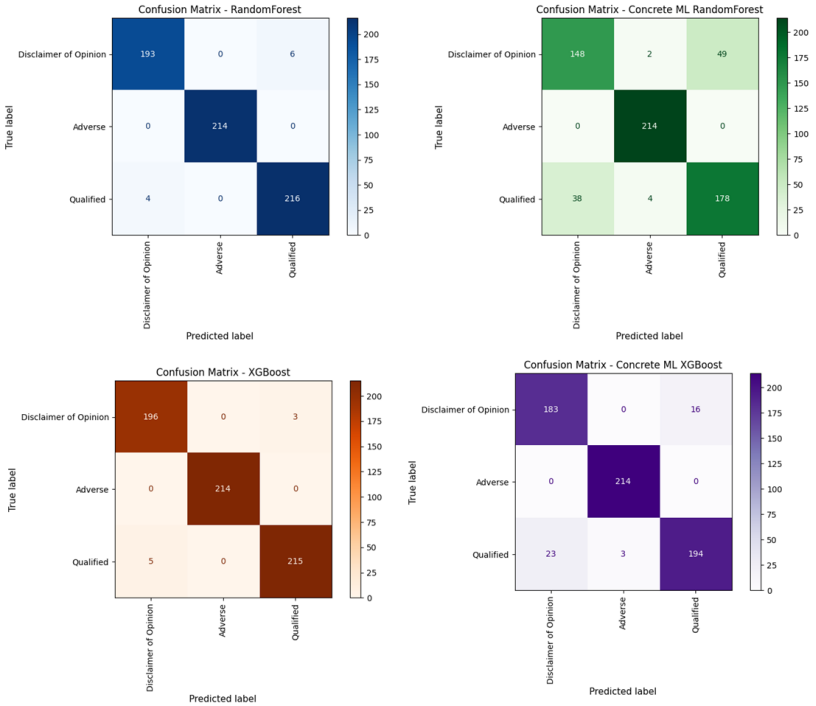

Figure 11. Confusion matrices obtained from Classifier-3 results

3.3.4 Classifier-4 Results: *Original Classes (Phase B)*

Phase B utilizes a simple flat classification framework aimed at predicting the four types of audit opinions. Figure 12 shows that the XGBoostClassifier has a slight edge over the rest of the models since it has the highest scores overall.

The RandomForestClassifier also has stable and accurate performance when compared on all the evaluation metrics. The Concrete ML models perform slightly less compared to their non-encrypted counterparts but remain well-performing models, and the Concrete ML RandomForest remains competitive on recall scores.

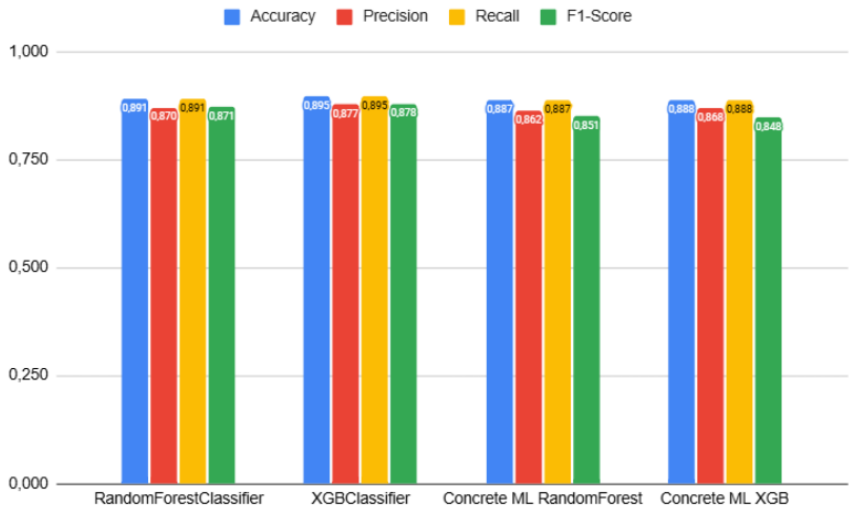

Figure 12. The results of the classification model, whose architecture is shown in Figure 3.

These results reflect the overall strength of traditional models in direct multi-class classification, while also showing that encrypted models remain practical alternatives in privacy-sensitive environments.
A detailed analysis of the models used in Phase B is presented by the confusion matrices shown in Figure-13.

As shown in Figure 3, the overall accuracy of the hierarchical classification model was calculated using confusion matrices. The accuracy results are as follows: Random Forest achieved 85.3%, XGBoost achieved 86.7%, Concrete ML XGBoost achieved 58.2%, and Concrete ML Random Forest achieved 67.0%. The classification performance in Phase A was higher than in Phase B.

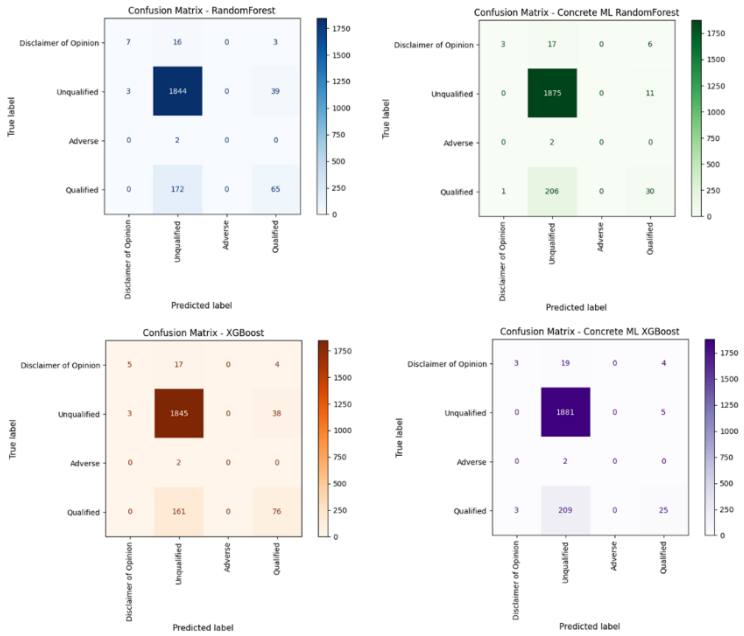

Figure 13. The confusion matrixes obtained from results of the classification model whose architecture is shown in Figure 3

4. CONCLUSION

This study presents a comparative analysis of hierarchical and classification architectures for the prediction of independent audit opinions. Audit opinions are examined both at the level of the four main categories (Unqualified, Qualified, Adverse, and Disclaimer of Opinion) and through the refinement of unqualified opinions into sub-categories. A novel risk score, derived from financial ratios, has been shown to significantly enhance the classification performance of the model.

The classification task was carried out using XGBoost and Random Forest algorithms, while class imbalance was mitigated using the SMOTE technique. In the flat classification approach (Phase B), the XGBoost algorithm achieved the highest performance, with an accuracy of 85.9% and an F1-score of 84.7%. In the hierarchical classification approach (Phase A), XGBoost achieved an accuracy of 89.6% and an F1-score of 88.1% in Classifier-1 (Unqualified vs. Others). For Classifier-2 (sub-classes of Unqualified), both XGBoost and Random Forest showed consistently strong performance, with accuracy levels exceeding 85%. In Classifier-3 (Qualified, Adverse, Disclaimer), the SMOTE-enhanced XGBoost model improved classification accuracy and reduced errors, particularly in minority classes.

Predictions made on encrypted data using Concrete ML yielded only marginal performance degradation. The encrypted version of XGBoost achieved 89.5% accuracy and 86.3% F1-score, demonstrating that high classification performance can be preserved while ensuring data confidentiality.

In conclusion, the proposed framework offers robust and reliable predictions across both conventional and privacy-preserving environments, contributing meaningfully to the audit process. It enables a more systematic, interpretable, and objective decision-support mechanism for auditors and stakeholders prior to the formal audit. Future work may extend this framework to different sectors to evaluate its generalizability, and may incorporate deep learning or explainable AI (XAI) techniques to enhance model interpretability further.

REFERENCES

[1] A. Yaşar, E. Yakut, M. M. Gutnu, “Predicting qualified audit opinions using financial ratios: Evidence from the Istanbul Stock Exchange”, *International Journal of Business and Social Science*, 6(8), 57–67, 2015.

[2] K. H. Chan, K. Z. Lin, R. R. Wang, “Government ownership, Accounting-Based regulations, and the pursuit of favorable audit opinions: Evidence from China”, Auditing: *A Journal of Practice & Theory*, 31(4), 47–64, 2012.

[3] N. Dopuch, R. W. Holthausen, R. W. Leftwich, “Predicting audit qualifications with financial and market variables”, *Accounting Review*, 62(3), 431–454, 1987.

[4] G. Husain, D. Nasef, R. Jose, J. Mayer, M. Bekbolatova, T. Devine, M. Toma, “SMOTE vs. SMOTEENN: A study on the performance of resampling algorithms for addressing class imbalance in regression models”, *Algorithms*, 18(1), 37, 2025.

[5] E. Kirkos, C. Spathis, A. Nanopoulos, Y. Manolopoulos, “Identifying Qualified Auditors’ Opinions: A Data Mining Approach”, *Journal of Emerging Technologies in Accounting*, 4, 183–197, 2007.

[6] P. T. T. Oanh, D. N. Hung, V. T. T. Van, “Forecasting Audit Opinions on Financial Statements: Statistical Algorithm or Machine Learning”, *Electronic Journal of Applied Statistical*

Analysis, 17(1), 133–152, 2024.

[7] A. Habib, "A meta-analysis of the determinants of modified audit opinion decisions", *Managerial Auditing Journal*, 28(3), 184–216, 2013.

[8] S. M. Saif, M. Sarikhani, F. Ebrahimi, "Finding Rules for Audit Opinions Prediction Through Data Mining Methods", *European Online Journal of Natural and Social Sciences*, 1(2), 28–36, 2012.

[9] J. R. Sánchez-Serrano, D. Alaminos, F. García-Lagos, "Predicting Audit Opinion in Consolidated Financial Statements with Artificial Neural Networks", *Mathematics*, 8, 1288, 2020.

[10] D. D. Tan, "Forecasting Audit Opinions on Financial Statements: Statistical Algorithm or Machine Learning", *Engineering and Technology Journal*, 2024.

[11] T. Chen, C. Guestrin, "XGBoost: A Scalable Tree Boosting System", **Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2016)**, San Francisco, CA, A.B.D., 785–794, 2016.

[12] Chang Yu, Yixin Jin, Qianwen Xing, Ye Zhang, Shaobo Guo, Shuchen Meng. **Advanced User Credit Risk Prediction Model using LightGBM, XGBoost and Tabnet with SMOTEENN.arXiv preprint arXiv:2408**.03497v3, 2024.

[13] L. Breiman, "Random Forests", *Machine Learning*, 45(1), 5–32, 2001.

[14] N. Stanišić, T. Radojević, N. Stanić, "Predicting the type of auditor opinion: Statistics, machine learning, or a combination of the two?", *Machine Learning*, 1–58, 2019.

[15] A. K. Nawaiseh, M. F. Abbod, T. Itagaki, "Financial Statement Audit using Support Vector Machines, Artificial Neural Networks and K-Nearest Neighbor: An Empirical Study of UK and Ireland", *International Journal of Simulation–Systems, Science & Technology*, 21(2), 2020.

[16] M. El-Bannany, M. Sreedharan, A. M. Khedr, "A robust deep learning model for financial distress prediction", *International Journal of Advanced Computer Science and Applications*, 11(2), 2020.

[17] M. Elhoseny, N. Metawa, G. Sztanó, I. M. El-Hasnony, "Deep learning-based model for financial distress prediction", *Annals of Operations Research*, 345(2), 885–907, 2025.

[18] E. I. Altman, M. Iwanicz-Drozdowska, E. K. Laitinen, A. Suvas, "Financial distress prediction in an international context: A review and empirical analysis of Altman's Z-score model", *Journal of International Financial Management & Accounting*, 28(2), 131–171, 2017.

[19] A. Yaşar, "Olumlu görüş dışındaki denetim görüşlerinin veri madenciliği yöntemleriyle tahminine ilişkin karar ve birliktelik kuralları", *Mali Çözüm Dergisi / Financial Analysis*, 26(133), 2016.

[20] H. Zarei, H. Yazdifar, M. D. Ghaleno, R. Azhmaneh, "Predicting auditors' opinions using financial ratios and non-financial metrics: Evidence from Iran", *Journal of Accounting in Emerging Economies*, 10(3), 425–446, 2020

[21] M. Brygała, T. Korol, "Personal bankruptcy prediction using machine learning techniques", *Economics and Business Review*, 10(2), 118–142, 2024.

[22] J. P. Sánchez Ballesta, E. García-Meca, "Audit qualifications and corporate governance in Spanish listed firms", *Managerial Auditing Journal*, 20(7), 725–738, 2005.

[23] G. Husain, D. Nasef, R. Jose, J. Mayer, M. Bekbolatova, T. Devine, M. Toma, "SMOTE vs. SMOTEENN: A study on the performance of resampling algorithms for addressing class imbalance in regression models", *Algorithms*, 18(1), 37, 2025.

[24] N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique", J. Artif. Intell. Res., 16, 321–357, 2002.

[25] B. Adiloğlu, B. Vuran, "A multicriterion decision support methodology for audit opinions: The case of audit reports of distressed firms in Turkey", Int. Bus. Econ. Res. J., 10(12), 37–48, 2011.

[26] A. Sideras, K. Bougiatiotis, E. Zavitsanos, G. Paliouras, G. Vouros, "Bankruptcy Prediction: Data Augmentation, LLMs and the Need for Auditor's Opinion", **Proceedings of the 5th ACM International Conference on AI in Finance, Association for Computing Machinery**, 453–460, November 2024.

[27] U. Gupta, GPT-InvestAR: Enhancing stock investment strategies through annual report analysis with large language models, arXiv preprint arXiv:2309.03079, 2023.

[28] Y. Huang, Z. Wang, C. Jiang, "Diagnosis with incomplete multi-view data: A variational deep financial distress prediction method", Technol. Forecast. Soc. Change, 201, 1–12, 2024.

[29] A. Saeedi, "A high-dimensional approach to predicting audit opinions", Applied Economics, 55(33), 3807–3832, 2023.

[30] H. Zarei, H. Yazdifar, M. D. Dahmarde Ghaleno, R. Azhmaneh, "Predicting auditors' opinions using financial ratios and non-financial metrics: evidence from Iran", *Journal of Accounting in Emerging Economies*, 10(3), 425–446, 2020.

[31] M. Todorovic, N. Stanisic, M. Zivkovic, N. Bacanin, V. Simic, E. B. Tirkolaee, "Improving audit opinion prediction accuracy using metaheuristics-tuned XGBoost algorithm with interpretable results through SHAP value analysis", Applied Soft Computing, 149, 110955, 2023.

[32] Agra Fintech, B. Aktürk, AgraResearchLab, T. Büyüktanır, "Audit opinions of Turkish Public Companies [Data Set]", Kaggle, https://www.kaggle.com/datasets/agrafintech/financial-data-of-turkish-public-companies, 28.05.2025.

[33] A. Stoian, B. Chevallier-Mames, "Zama Concrete ML: Simplifying Homomorphic Encryption for Python Machine Learning", Python.org, https://www.python.org/success-stories/zama-concrete-ml-simplifying-homomorphic-encryption-for-python-machine-learning/, 28.05.2025.

[34] F. T. Kristanti, M. Y. Febrianta, D. F. Salim, H. A. Riyadh, Y. Sagama, B. A. H. Beshr, "Advancing Financial Analytics: Integrating XGBoost, LSTM, and Random Forest Algorithms for Precision Forecasting of Corporate Financial Distress", *Journal of Infrastructure, Policy and Development*, 8(8), 4972, 2024.

[35] T. K. Ho, "Random decision forests", **Proceedings of the 3rd International Conference on Document Analysis and Recognition, Lausanne, Switzerland**, 1, 278–282, IEEE, New York, NY, USA, 1995.

[36] K. Tissaoui, T. Zaghdoudi, A. Hakimi, M. Nsaibi, "Do gas price and uncertainty indices forecast crude oil prices? Fresh evidence through XGBoost modeling", *Computational Economics*, 62, 663–687, 2022.

[37] M. Imani, A. Beikmohammadi, H. R. Arabnia, "Comprehensive analysis of random Forest and XGBoost performance with SMOTE, ADASYN, and GNUS under varying imbalance levels", *Technologies*, 13(3), 88, 2025.

[38] E. Yılmaz, T. Büyüktanır, D. Civelek, B. Aktürk, AgraResearchLab, "Refined audit opinions using bankruptcy algorithms" [Data Set], Kaggle, https://www.kaggle.com/datasets/ensryilmaz/refined-audit-opinions-using-bankruptcy-algorithms/data, 28.05.2025

[39] F. T. Kristanti ve V. Dhaniswara, The accuracy of artificial neural networks and logit models in predicting the companies' *financial distress, Journal of Technology Management and Innovation*, 2023.

[40] S. B. Kotsiantis, Supervised machine learning: A review of classification techniques, Informatica, 2007.

[41] D. M. W. Powers, "Evaluation: From precision, recall and F-factor to ROC, informedness, markedness & correlation", *Journal of Machine Learning Technologies*, 2011, 2,