IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 7 December 2023, accepted 10 January 2024, date of publication 15 January 2024, date of current version 22 January 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3354173

RESEARCH ARTICLE

Advancing Bankruptcy Forecasting With Hybrid Machine Learning Techniques: Insights From an Unbalanced Polish Dataset

UMMEY HANY AINAN[1], LIP YEE POR[1], (Senior Member, IEEE), YEN-LIN CHEN[2], (Senior Member, IEEE), JING YANG[1], (Graduate Student Member, IEEE), AND CHIN SOON KU[3]

[1]Department of Computer System and Technology, Faculty of Computer Science and Information Technology, Universiti Malaya, Kuala Lumpur 50603, Malaysia
[2]Department of Computer Science and Information Engineering, National Taipei University of Technology, Taipei 106344, Taiwan
[3]Department of Computer Science, Universiti Tunku Abdul Rahman, Kampar 31900, Malaysia

Corresponding authors: Yen-Lin Chen (ylchen@mail.ntut.edu.tw), Lip Yee Por (porlip@um.edu.my), and Chin Soon Ku (kucs@utar.edu.my)

This work was supported in part by the National Science and Technology Council in Taiwan under Grant NSTC-112-2221-E-027-088-MY2 and Grant NSTC-111-2622-8-027-009; in part by the Ministry of Education of Taiwan through "The Study of Artificial Intelligence and Advanced Semiconductor Manufacturing for Female Science, Technology, Engineering, and Mathematics (STEM) Talent Education and Industry-University Value-Added Cooperation Promotion" under Grant 1122302319; and in part by the Universiti Tunku Abdul Rahman (UTAR) Financial Support for Journal Paper Publication Scheme through UTAR, Malaysia.

ABSTRACT The challenge of bankruptcy prediction, critical for averting financial sector losses, is amplified by the prevalence of imbalanced datasets, which often skew prediction models. Addressing this, our study introduces the innovative hybrid model XGBoost+ANN, designed to leverage the strengths of both ensemble learning and artificial neural networks. This model integrates a comprehensive set of features with parameters optimized through genetic algorithms, eschewing traditional feature selection approaches. Our research focuses on an unbalanced dataset of Polish companies and reveals that the XGBoost+ANN model, in particular, exhibits outstanding performance. Optimized using genetic algorithms and without feature selection, this model achieved the highest AUC (0.958), sensitivity (0.752), and accuracy (0.983) scores, surpassing other models in our study. This remarkable outperformance, along with the robust results, marks a substantial advancement in the field of bankruptcy prediction. It underscores the efficacy of our approach in addressing the persistent challenge of data imbalance, offering a more reliable and accurate solution for financial risk assessment.

INDEX TERMS Bankruptcy forecasting, predictive analytics, ensemble learning, hyperparameter tuning, machine learning.

I. INTRODUCTION

The prediction of bankruptcy is a crucial concern in the financial sector, with significant implications for business stability and economic health. Accurate predictions of financial distress are essential, as company bankruptcies can lead to job losses, supply chain disruptions, reduced tax revenue, and shaken investor confidence, all of which contribute to economic instability and hinder growth.

Numerous methods for predicting firm bankruptcy have been proposed, including classical statistical methods, machine learning, and artificial intelligence [1], [2], [3]. These methods have been applied to both balanced datasets, where the numbers of failed and successful companies are nearly equal [4], [5], and to balanced datasets created through sampling techniques such as under-sampling and over-sampling (SMOTE) [3], [6]. However, it is nearly impossible

The associate editor coordinating the review of this manuscript and approving it for publication was Ehab Elsayed Elattar.

to have a balanced bankruptcy dataset in the real world, leading researchers to work with unbalanced datasets where instances of bankruptcy are far outnumbered by non-bankrupt cases. To address this imbalance, studies have applied various models—hybrid, statistical, individual, and ensemble machine learning—to such datasets [7], [8]. Although some of these methods have not produced satisfactory results and require further improvement, it is critical to adequately train models on all classes of data to ensure accuracy during testing.

Given the explosion of data and the quest for more precise results, hybrid machine learning methods are increasingly popular. We propose the use of the hybrid method XGBoost+ANN. These machine learning algorithms, capable of processing vast quantities of data, aid in making superior business decisions. Our approach also includes individual and ensemble machine learning methods (SVM, RF, and XGBoost), as well as feature selection and optimization techniques, which further enhance results. Selecting the right metrics to evaluate performance on skewed datasets is vital [9], and our review suggests that AUC is the most appropriate measure.

The paper is organized as follows: Section II reviews relevant literature; Section III details the research methodology and describes the proposed method configurations; and Section IV presents and discusses results using a publicly available Polish company dataset. Using public data allows us to compare our approach against others previously published, ensures the reproducibility of our results, and demonstrates that our proposed hybrid method outperforms existing ones. Section V concludes by summarizing the research's strengths and weaknesses and suggesting directions for future research.

II. RELATED WORK

Bankruptcy prediction, a crucial field within finance and risk management, has developed in tandem with advancements in data analysis, statistical methods, and modern computational tools. Originally, simple statistical techniques were used to analyze financial ratios to identify patterns indicative of bankruptcy risk [3]. A significant development was Edward Altman's Z-Score model, which combined multiple financial ratios into a single predictive score [3]. While statistical methods were once dominant in bankruptcy prediction, the emergence of machine learning has caused a shift, offering advantages over traditional statistical approaches [3], [4], [10], [11].

Various single machine learning models, such as logistic regression (LR), support vector machines (SVM), decision trees (DT), and rule-based models like jRip and J48, have been applied to both balanced and unbalanced datasets for bankruptcy prediction, with DTs noted for their consistent performance across dataset types [12]. However, models like LR and SVM can struggle with unbalanced datasets. Ensemble methods, including Random Forest, AdaBoost, Gradient Boosting, XGBoost, and CatBoost, have improved predictive performance, with Random Forest and XGBoost, in particular, being known for their accuracy and robustness in handling diverse datasets [13], [14], [15], [16].

Multilayer perceptron, a type of artificial neural network (ANN), has also gained attention in bankruptcy prediction [17], [18], [19], [20]. Recent trends include the exploration of hybrid models that combine different machine-learning algorithms [21] and domain knowledge to improve accuracy and robustness. Notable among these are the HAOC, which integrates oversampling frameworks with the cBoost algorithm [6], and a two-stage hybrid learning approach that combines statistical and machine learning clustering with classification [22].

Feature selection and optimization techniques are equally important in refining predictive models. These methods range from model-free algorithms to gradient-based and Bayesian optimization. Metaheuristic algorithms, such as genetic algorithms (GA) and particle swarm optimization (PSO), often provide effective solutions for complex problems [23], [24]. GAs are particularly significant in efficiently identifying relevant features.

A class imbalance in financial datasets, typically biased towards solvent companies, presents a major challenge. It often leads to the misclassification of the minority class [6], [8], [12], [15], [19], [25], [26], [27]. To address this, class-balancing methods like under-sampling, oversampling (including SMOTE and its variants), and combined techniques are frequently used. Some studies also explore unbalanced datasets without prior class balancing [7], [8], [13], [24]. The ongoing development in bankruptcy prediction reflects the continuous pursuit of more accurate, efficient, and reliable methods to anticipate financial distress and mitigate its effects.

Despite these advancements, challenges persist due to the evolving nature of financial data and inherent class imbalances in datasets. This necessitates the development of more robust and precise methodologies. The dynamic field of bankruptcy prediction demands innovative approaches that can adapt to diverse datasets, manage class imbalances, and enhance the accuracy and efficiency of predictive models, ensuring reliable predictions of financial distress and its mitigation. To address these challenges, the following method is proposed in the next section:

III. PROPOSED METHOD

The proposed method encompasses a five-stage process outlined in Figure 1, starting with data collection and followed by rigorous data preprocessing. Parameter optimization is then performed to fine-tune the models for peak efficiency. Subsequently, feature selection techniques are applied to determine the most relevant features. Finally, the proposed model is implemented along with existing classifiers to enable accurate predictions.

A. DATASET

The dataset utilized in this study encompasses data on the bankruptcy of Polish companies. This dataset is well-known

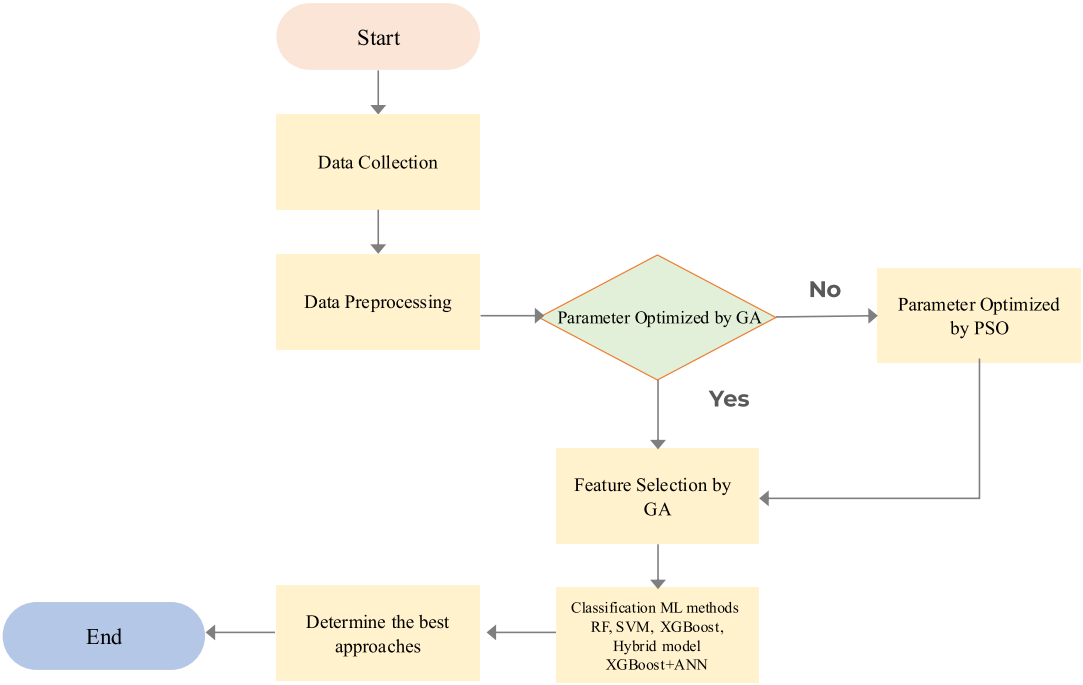


FIGURE 1. Flowchart of the proposed method.

TABLE 1. Polish company dataset description.

Data	Records
Number of attributes	64
Number of instances	7027
Successful companies	6756
Bankrupted companies	271
percentage	96% successful and 4% bankrupt
Source	UCI

and extensively utilized in financial distress prediction research [9], [13], [24], [28]. The data, which pertains to the probability of bankruptcy for Polish businesses, was sourced from the Emerging Markets Information Service (EMIS) database, providing details on global emerging markets. It consists of 7027 instances and 64 attributes and is accessible from the UCI Machine Learning Repository (https://doi.org/10.24432/C5F600). Table 1 contains detailed information about the dataset.

B. DATA PREPROCESSING

Data preprocessing is of critical importance in data science and machine learning. It ensures the data is clean, standardized, and ready for use, thereby enabling machine learning models to make accurate predictions, uncover valuable insights, and support informed decision-making across various applications. The data preprocessing steps undertaken in this study are as follows:

- Data Cleaning: Missing values are addressed through mean imputation, which involves replacing the missing values in each column with the column's mean. This step is crucial to ensuring the completeness and cleanliness of the data and resolving any issues with missing or unknown values.
- Normalization: The data is normalized using the Min-Max normalization technique (see equation 1), which scales all numerical features to a range of [0, 1], providing consistency in scale [29]. Min-Max normalization is critical for machine learning algorithms, such as SVM and ANN, that depend on distance calculations or gradient-based optimization, as it prevents certain features from overpowering others due to scale differences.

$$X' = \frac{X - \min(X)}{\max(X) - \min(X)} \tag{1}$$

where X' represents the normalized value, X represents the actual value, and $\min(X)$ and $\max(X)$ represent the minimum and maximum values of X, respectively [30].

C. PARAMETER OPTIMIZATION

Parameter optimization is a critical step in creating an effective machine-learning model. This process involves fine-tuning hyperparameters to optimize model performance. Hyperparameters, which dictate the behavior of a learning algorithm, are crucial for the success of machine learning methods with numerous hyperparameters, such as Random Forest (RF), XGBoost, Support Vector Machines (SVM), and Artificial Neural Networks (ANN). The process of finding optimal hyperparameters is known as parameter optimization [23]. Metaheuristic optimization methods, inspired by natural processes, are frequently used to efficiently navigate large parameter spaces. This study employs genetic algorithms and PSO for parameter optimization.

Genetic algorithms are chosen for their robustness and proven effectiveness in exploring large and complex search spaces, mimicking the process of natural evolution to efficiently navigate towards optimal solutions. They excel in handling both discrete and continuous variables, making them versatile for various parameter optimization tasks. Particle swarm optimization is used for its simplicity and ability to converge quickly to a near-optimal solution. Inspired by the social behavior patterns of animals, PSO is particularly effective in multidimensional search spaces, where it leverages collective and individual learning processes to guide the search, thus complementing the evolutionary search approach of genetic algorithms.

1) GENETIC ALGORITHM (GA)

The Genetic Algorithm (GA), a method of evolutionary computation, is based on the principles of natural selection [31]. In GAs, individuals best suited to their environment are more likely to survive and pass on their traits to subsequent generations. As generations progress, the population evolves, comprising both superior and inferior individuals who inherit characteristics from their predecessors. Over time, less fit individuals are naturally eliminated, while the fitter ones survive, producing more capable offspring. The most adaptable individual is eventually identified as the global optimum after many generations [23]. GA is renowned for its efficient search capabilities in large solution spaces, adaptability across various domains, and resilience to noise and local optima, making it a powerful tool for parameter optimization in machine learning.

The efficiency of GA in searching for and selecting the most suitable combinations of parameters or features for a given problem by emulating natural selection and evolution processes is the reason for its use in this study for both parameter optimization and feature selection.

2) PARTICLE SWARM OPTIMIZATION (PSO)

PSO is a population-based optimization algorithm inspired by the social behavior of animals, such as birds and fish. It seeks the optimal solution by effectively exploring the search space [37]. The population, or swarm, comprises individuals known as particles, each defined by a specific position and velocity.

The PSO process involves iterative improvement of solutions. Each particle updates its position based on its best experience (pbest) and the overall best swarm experience (gbest). The swarm's collective knowledge and individual experiences guide the particles in exploring promising regions of the search space. Position and velocity updates occur at each iteration, as defined by the following equations [32]:

$$\begin{aligned} &\textit{Velocity Equation} : V_m(i+1) \\ &\quad = w(i) \times V_m(i) + c1 \times r1(pbest_m(i)x_m(i)) \\ &\quad + c2 \times r2(gbest(i)x_m(i)) \end{aligned} \tag{2}$$

$$\begin{aligned} &\textit{Position Equation} : X_m(i+1) \\ &\quad = V_m(i+1) + x_m(i) \end{aligned} \tag{3}$$

Here, x_{m} and V_{m} represent the particle's position and velocity, respectively. m is the particle index and i denotes the iteration number. The inertia weight is w, while c_1 and c_2 are learning factors. r_1 and r_2 are random values between 0 and 1, and Num_{p} and Num_{i} denote the total number of particles and the maximum number of iterations, respectively.

The fitness function evaluates the performance of each particle's position, guiding the optimization of hyperparameter values for a specific task. PSO continues through iterations until either a predetermined number is completed or a specific fitness level is achieved, according to the chosen termination criteria.

Given its ability to efficiently explore multidimensional search spaces and facilitate the discovery of near-optimal solutions through dynamic particle interaction, PSO is used in this research for parameter optimization.

D. FEATURE SELECTION

A typical real-world dataset may possess numerous features. Feature selection is the process of selecting a subset of pertinent features from a larger pool, aiming to reduce data dimensionality by retaining only relevant features and discarding the irrelevant or redundant ones [33]. This study employs genetic algorithms (GA) for feature selection, a method that is part of the random feature selection category (see Figure 2).

1) GENETIC ALGORITHM (GA)

GA, a pioneering population-based stochastic algorithm, operates on selection, crossover, and mutation processes [34]. The components of GA include [31]:

- Population: GA begins with an initial population of potential solutions (feature subsets), each represented by a binary string, where each bit signifies the inclusion or exclusion of a feature.
- Fitness Function: This function evaluates the quality of each solution (feature subset) by quantifying its performance in the machine learning task, thereby determining the likelihood of each solution's selection for reproduction.
- Selection: Solutions are chosen from the population based on their fitness scores, with higher fitness increasing the chances of selection.
- Crossover: Selected feature subsets undergo genetic recombination to produce new candidate solutions (offspring), promoting diverse and potentially improved feature combinations.
- Mutation: Random changes are introduced to some individuals, adding diversity and enabling the exploration of new feature combinations.

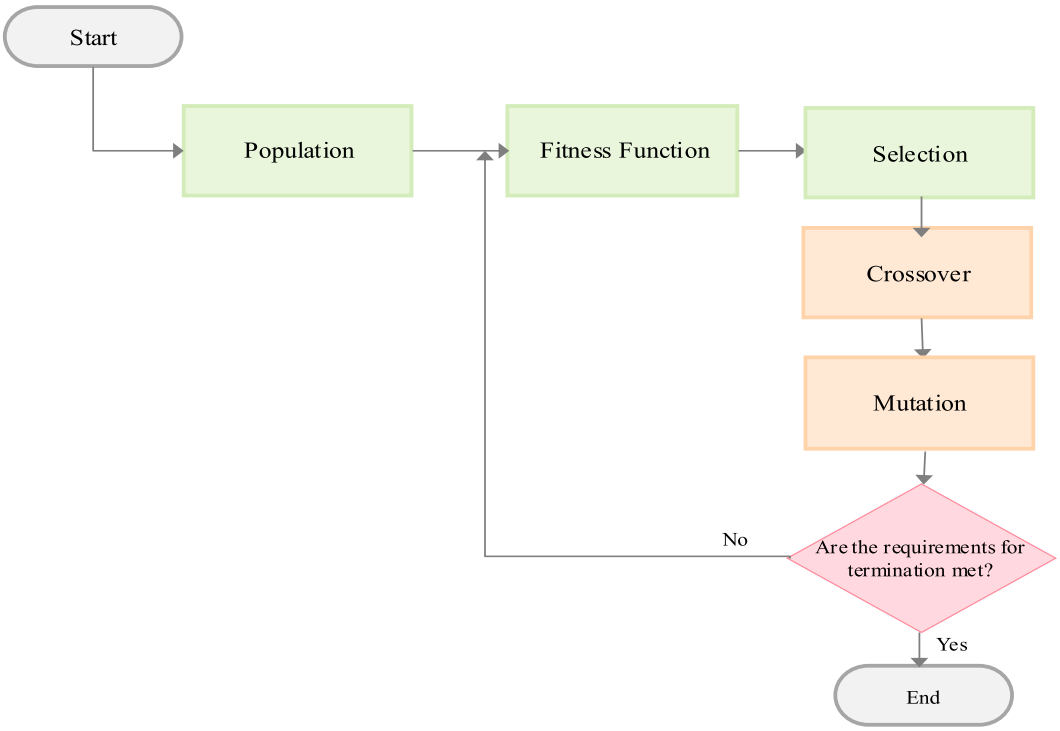


FIGURE 2. Genetic algorithm [34].

- Terminate: The algorithm concludes after a set number of generations or upon meeting a fitness-related stopping criterion.

GA seeks the most informative feature subset, iterating across generations to refine the set and enhance model accuracy.

E. CLASSIFICATION METHODS

This study employs three classification methods—Random Forest (RF), Extreme Gradient Boosting (XGBoost), and Support Vector Machine (SVM)—selected for their prevalent use in bankruptcy prediction. Random Forest's ensemble approach, which combines multiple decision trees, provides robustness against class imbalance, a common challenge in bankruptcy datasets. This robustness stems from its ability to reduce variance and overfitting, making the model more generalizable to diverse data scenarios. XGBoost's boosting technique is adept at adapting to minority class patterns, which is crucial in bankruptcy prediction where default cases are often less represented. It achieves this by sequentially focusing on and correcting misclassified instances from previous iterations, thus enhancing the overall predictive power. SVM, known for its effectiveness with high-dimensional data, can effectively create decision boundaries through the use of appropriate kernel functions. These kernels transform the data into a higher dimension where it becomes easier to segregate classes linearly, making SVM particularly useful for complex patterns often encountered in financial data. Collectively, these methods enhance predictive accuracy in the context of imbalanced bankruptcy data by leveraging their individual strengths—RF's variance reduction, XGBoost's focus on misclassifications, and SVM's kernel trick—to address the unique challenges of bankruptcy prediction.

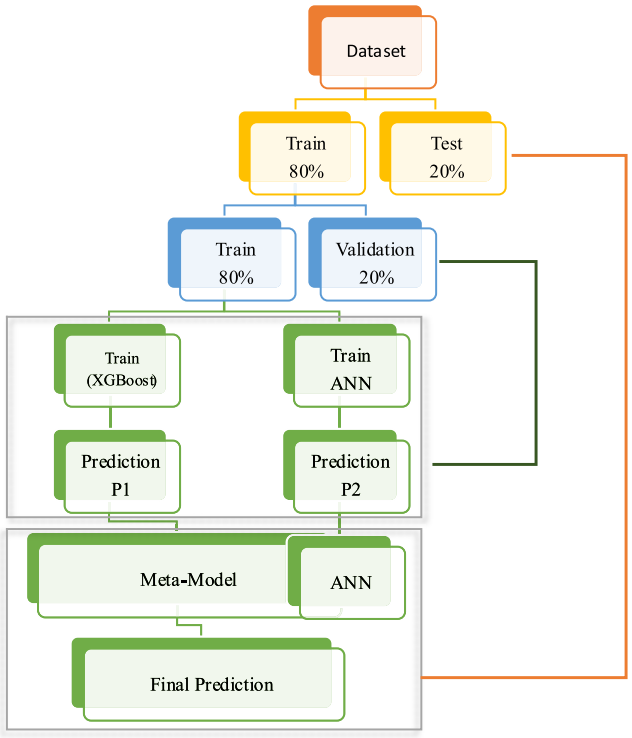


FIGURE 3. Proposed hybrid methods.

TABLE 2. The selected features were found by the genetic algorithm.

Dataset Variable ID	Selected Feature by GA
X3	working capital / total assets
X4	current assets / short-term liabilities
X5	(cash + short-term securities + receivables - short-term liabilities) / (operating expenses - depreciation) * 365
X6	retained earnings / total assets
X7	EBIT / total assets
X8	book value of equity / total liabilities
X9	sales / total assets
X11	(gross profit + extraordinary items + financial expenses) / total assets
X16	(gross profit + depreciation) / total liabilities
X18	gross profit / total assets
X21	sales (n) / sales (n-1)
X22	profit on operating activities / total assets
X24	gross profit (in 3 years) / total assets
X25	(equity - share capital) / total assets
X26	(net profit + depreciation) / total liabilities
X27	profit on operating activities / financial expenses
X29	logarithm of total assets
X30	(total liabilities - cash) / sales
X33	operating expenses / short-term liabilities
X34	operating expenses / total liabilities
X37	(current assets - inventories) / long-term liabilities
X39	profit on sales / sales
X43	rotation receivables + inventory turnover in days
X45	net profit / inventory
X46	(current assets - inventory) / short-term liabilities
X47	(inventory * 365) / cost of products sold
X51	short-term liabilities / total assets
X52	(short-term liabilities * 365) / cost of products sold)
X54	constant capital / fixed assets
X57	(current assets - inventory - short-term liabilities) / (sales - gross profit - depreciation)
X58	total costs /total sales
X59	long-term liabilities / equity
X62	(short-term liabilities *365) / sales

Note: Dataset Variable ID: Polish company bankruptcy data variable ID from the UCI machine learning repository (https://doi.org/10.24432/C5F600).

1) RANDOM FOREST (RF)

Introduced in 2001, RF is an ensemble learning method that combines the predictions of multiple decision trees. It uses bootstrapping to train each tree on different data subsets, with the final prediction determined by the majority vote of these trees [35]. RF is renowned for its accuracy, efficiency, robust handling of missing data, and ability to process extensive feature sets [36].

2) EXTREME GRADIENT BOOSTING (XGBOOST)

XGBoost is a gradient-boosting framework designed to improve the performance of weak learners through iterative enhancements. It optimizes an objective function using gradient descent techniques. XGBoost is notable for its accuracy, interpretability, and effectiveness in handling missing data and class imbalances [10].

3) SUPPORT VECTOR MACHINE (SVM)

SVM operates by searching for the optimal hyperplane to maximize the margin between different classes. The support vectors are the data points that are nearest to this hyperplane. SVM excels at managing high-dimensional data and is particularly robust against overfitting. This makes it suitable for handling both linear and non-linear data sets [37].

F. PROPOSED HYBRID METHOD

The proposed hybrid model, XGBoost+ANN, integrate multiple models to improve accuracy and robustness (see Figure 3). We use the meta-learning concept, with ANN serving as the meta-model that synthesizes predictions from the base models (XGBoost and ANN).

The dataset is split 80/20 for training and testing, with further partitioning for validation. Base models are trained on the training data and evaluated on the validation data. The meta-model is trained on the predicted probabilities from the base models and tested on the original test data to gauge performance.

G. EVALUATION METRICS

To gauge the predictive efficacy of the models in this study, various performance metrics are employed, each chosen for its specific relevance to the context of bankruptcy prediction. Five key metrics have been selected: accuracy, precision, AUC (Area Under the Curve), sensitivity, and specificity.

1) ACCURACY

Accuracy is a prevalent metric in classification tasks. It measures the ratio of correctly predicted instances (true positives and true negatives) to the total dataset instances, essentially reflecting the model's proficiency in correctly

TABLE 3. The optimal hyperparameter found by the GA.

Method	Parameters	Search Space	Optimal parameter value by GA
RF	n_estimators	[1,100]	100
	max_depth	[1,20]	1
SVM	Kernel	['linear', 'poly', 'rbf']	linear
	C	[0.1, 1.0, 10.0]	10.0
	gamma	[0.01, 0.1, 1.0]	0.1
XGBoost	learning_rate	[0.1, 0.01, 0.001]	0.1
	gamma	[0, 1, 5]	1
ANN	hidden_layer_sizes	[(50,), (100,), (50, 70), (100, 70), (100, 100)]	(50,)
	activation	['logistic', 'tanh', 'relu']	relu
	learning_rate	['constant', 'invscaling', 'adaptive']	adaptive

TABLE 4. The optimal hyperparameter found by PSO.

Method	Parameters	Search Space	Optimal parameter value by PSO
RF	n_estimators	[10,100]	40
	max_depth	[1,20]	14
SVM	Kernel	['linear', 'rbf']	linear
	C	[0.1, 10]	9.825
	gamma	[0.001, 0.1]	0.084
XGBoost	learning_rate	[0.01, 1.0]	1.0
	gamma	[0.001, 1.0]	0.001
ANN	hidden_layer_sizes	[10, 100]	38
	activation	['logistic', 'tanh', 'relu']	tanh
	learning_rate	['constant', 'invscaling', 'adaptive']	constant

TABLE 5. Comparative results between the proposed hybrid method and other methods with all features.

With all Features										
	Parameter optimized by GA					Parameter optimized by PSO				
Methods	AUC	Precision	Accuracy	Sensitivity	Specificity	AUC	Precision	Accuracy	Sensitivity	Specificity
RF	0.658	0.729	0.969	0.420	0.988	0.687	**0.942**	0.975	0.468	0.995
SVM	0.505	0.300	0.961	0.192	0.864	0.505	0.250	0.960	0.285	0.802
XGBoost	0.768	**0.950**	0.981	0.553	**0.998**	0.779	0.885	0.980	0.563	0.988
XGBoost+ANN	**0.958**	0.920	**0.983**	**0.752**	0.967	**0.939**	0.905	**0.982**	**0.656**	**0.996**

TABLE 6. Comparative results between the proposed hybrid method and other methods with selected features.

With selected Features										
	Parameter optimized by GA					Parameter optimized by PSO				
Methods	AUC	Precision	Accuracy	Sensitivity	Specificity	AUC	Precision	Accuracy	Sensitivity	Specificity
RF	0.715	**0.956**	0.977	0.429	0.998	0.711	**0.959**	0.977	0.475	0.992
SVM	0.504	0.200	0.961	0.115	0.917	0.503	0.150	0.961	0.285	0.802
XGBoost	0.776	0.948	**0.982**	0.542	0.985	0.761	0.822	0.978	0.552	0.996
XGBoost+ANN	**0.953**	0.923	0.981	**0.712**	**0.999**	**0.936**	0.874	**0.981**	**0.627**	**0.997**

classifying data points. The accuracy formula, as provided by [9], is:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{4}$$

Here, TP represents True Positives, TN represents True Negatives, FP stands for False Positives, and FN denotes False Negatives. Accuracy is important as it provides a straightforward measure of the model's overall effectiveness, covering all classifications.

2) PRECISION

Precision is crucial in binary classification to assess the model's capability to minimize false positives. It calculates the fraction of true positive predictions over all positive predictions made by the model, as shown in Equation 5 [9]:

$$Precision = \frac{TP}{TP + FP} \tag{5}$$

Precision is particularly significant in bankruptcy prediction, where the cost of false positives (predicting bankruptcy incorrectly) is high.

3) AREA UNDER THE CURVE (AUC)

AUC, or Area Under the Curve, is a key performance indicator in binary classification. The Receiver Operating Characteristic (ROC) curve plots the binary classifier's performance across different thresholds. AUC, representing

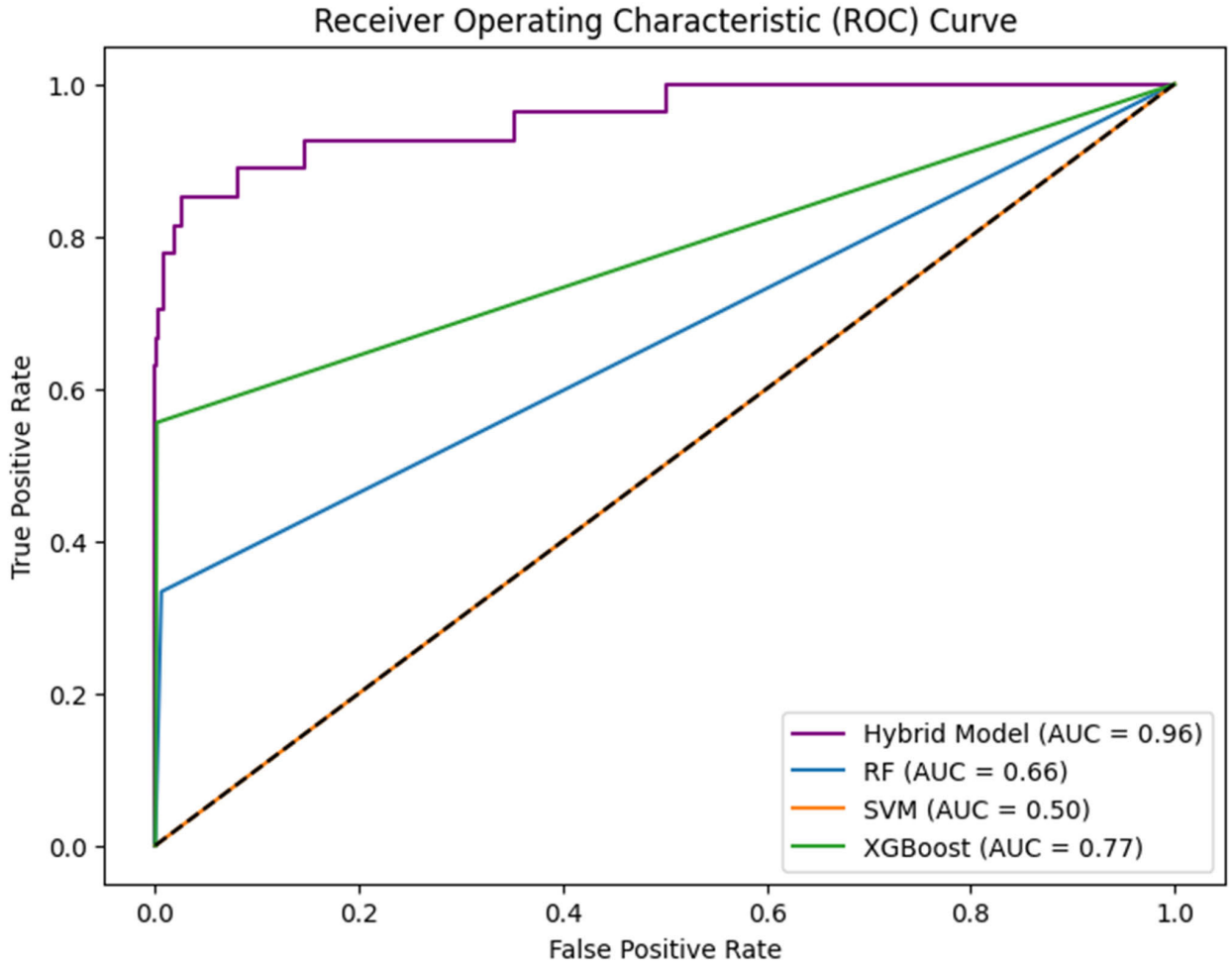


FIGURE 4. RF, SVM, XGBoost, and XGBoost+ANN with all features and parameters optimized by GA.

the area beneath the ROC curve, provides a singular value reflecting the model's aptitude for distinguishing between positive and negative classes. AUC values range from 0 to 1, with values closer to 1 indicating excellent model performance and those closer to 0 indicating poor performance.

Specifically, values within the range of 0.9 to 1 are considered exceptional, values from 0.7 to 0.9 are deemed good, and values below 0.7 are indicative of inadequate performance [25]. The formula for AUC, as provided by [38], is:

$$AUC = \frac{S_0 - n_0(n_0 + 1)/2}{n_0 n_1} \tag{6}$$

Here, n_0 and n_1 are the numbers of positive and negative examples, respectively, and $S_0 = \sum r_i$, where r_i is the rank of the i^{th} positive example in the ranked list. AUC is beneficial for imbalanced datasets as it accounts for the true positive and false positive rates across varying thresholds, thus providing a more nuanced metric for unbalanced data scenarios. It appraises the model's ranking efficacy for instances irrespective of class distribution, making it a more comprehensive measure in such contexts.

4) SENSITIVITY

Sensitivity, also called true positive rate or recall, measures the proportion of actual positive samples correctly predicted by the model. It's calculated as the ratio of True Positives (TP) to the sum of True Positives (TP) and False Negatives (FN), as shown in Equation 7 [24]. Sensitivity is crucial for identifying companies at risk early on and capturing the most actual positive instances.

$$Sensitivity = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{7}$$

5) SPECIFICITY

Specificity measures the proportion of actual negatives correctly identified by the model. It's calculated as the ratio of True Negatives (TN) to the sum of True Negatives (TN) and False Positives (FP), as shown in Equation 8 [24]. A high specificity indicates that the model is very good at accurately distinguishing non-bankrupt companies. In the context of bankruptcy prediction, high specificity is desirable as it minimizes the misclassification of financially stable

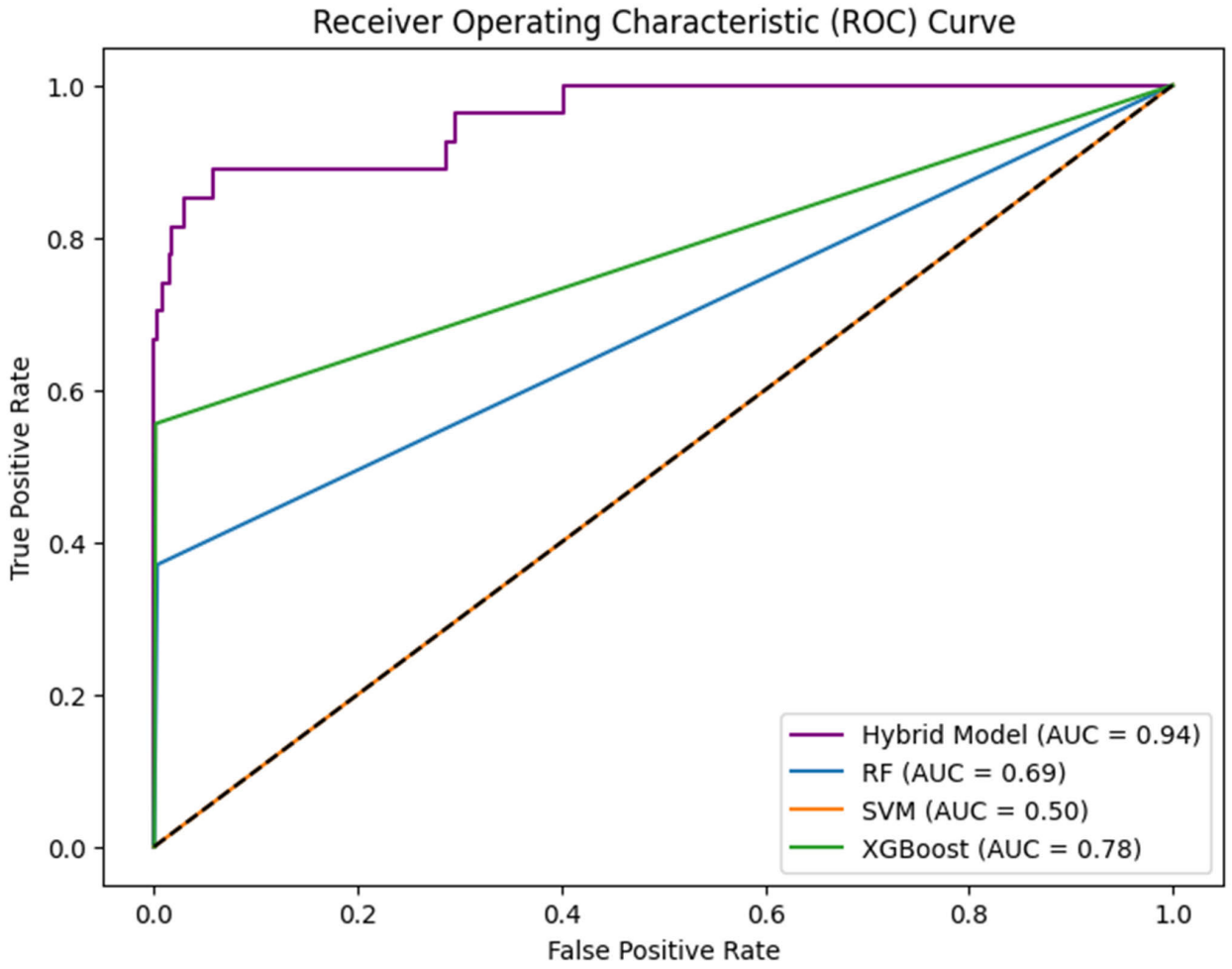


FIGURE 5. RF, SVM, XGBoost, and XGBoost+ANN with all features and parameters optimized by PSO.

companies as bankrupt.

$$Specificity = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{8}$$

IV. RESULTS AND DISCUSSION

The dataset under study was divided into training and testing sets, with an 80/20 split. Stratified 10-fold cross-validation was used to evaluate the model's performance, ensuring a consistent class distribution across folds—beneficial for imbalanced datasets.

A. SELECTED FEATURE

A genetic algorithm narrowed down the feature set from 64 to 33, as shown in Table 2.

This subset was determined by exploring various feature combinations over multiple generations and optimizing the classification methods used.

B. RESULT OF PARAMETER OPTIMIZATION

Different search spaces were defined for each parameter and optimizer, with GA and PSO being the chosen optimization methods. The optimal parameter values identified are listed in Tables 3 and 4. For the Random Forest model, the 'n_estimators' and 'max_depth' hyperparameters were found to be particularly influential. In the case of the = SVM model, the 'C', 'kernel', and 'gamma' parameters played a significant role. For the Extreme Gradient Boosting (XGBoost) model, the 'learning_rate' and 'gamma' parameters were fine-tuned, a process that proved to be especially beneficial for handling unbalanced datasets.

C. METHODS PERFORMANCE ANALYSIS

The performance of our proposed hybrid method was compared with traditional classification methods under two scenarios, with the best results highlighted in bold in Tables 5 and 6.

In Table 5, the XGBoost+ANN hybrid model, with its full range of features, demonstrates outstanding performance in terms of AUC, accuracy, and sensitivity. After careful parameter optimization using Genetic Algorithms (GA), this hybrid model achieves an AUC of 0.958, an accuracy of 0.983, and a sensitivity of 0.752. When Particle Swarm Optimization (PSO) is employed for parameter tuning, the model maintains a strong AUC of 0.939, with accuracy

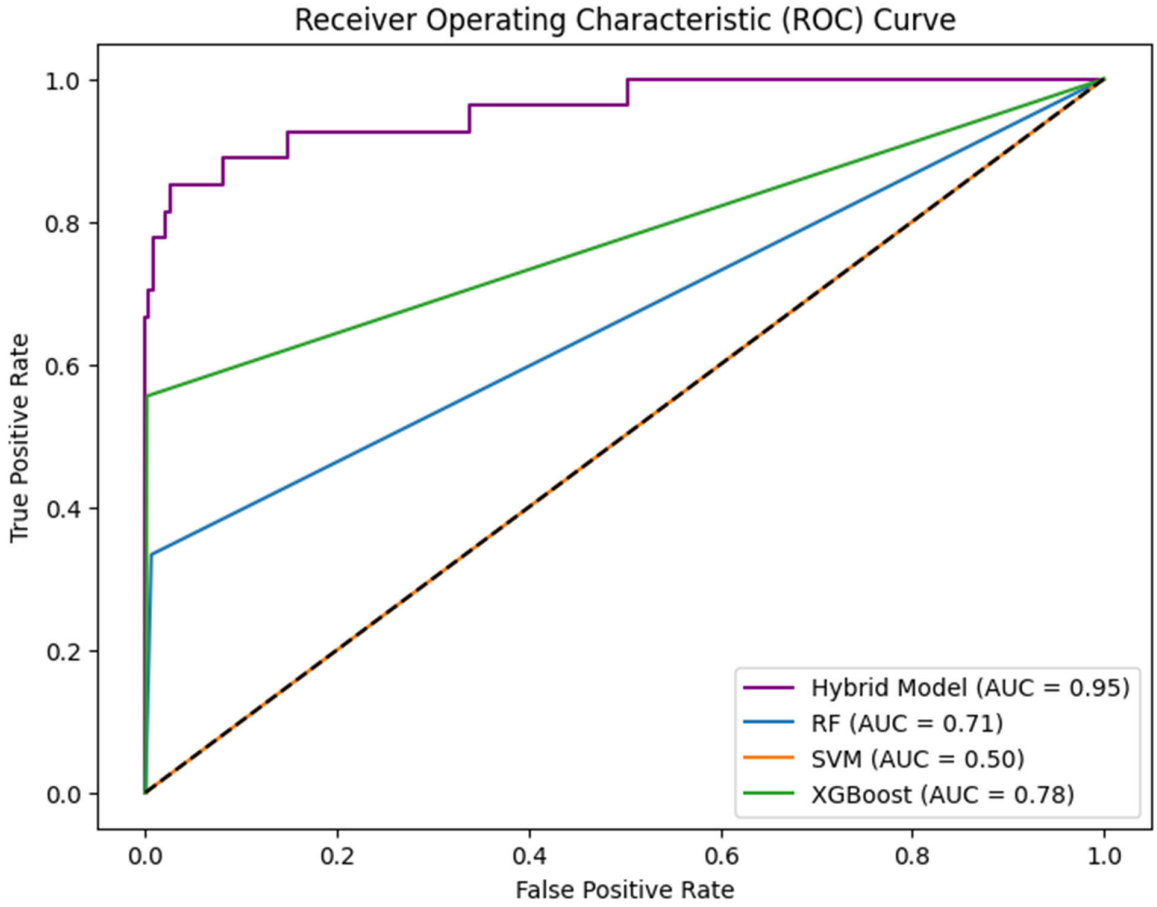


FIGURE 6. RF, SVM, XGBoost, and XGBoost+ANN with selected features and parameters optimized by GA.

and sensitivity slightly higher at 0.982 and 0.656. The XGBoost+ANN hybrid model also provides a higher specificity of 0.996.

In contrast, XGBoost excels in precision and specificity, reaching peaks of 0.950 and 0.998 when optimized with GA. Additionally, the RF model exhibits commendable precision (0.942) when optimized with PSO.

Table 6 continues to highlight the impressive performance of the XGBoost+ANN hybrid model, particularly in AUC, sensitivity, and specificity, with a score of 0.953, 0.712, and 0.999. When GA is used for parameter optimization, XGBoost achieves the highest accuracy of 0.982. The XGBoost+ANN with PSO reflects an admirable AUC of 0.936, closely followed by an accuracy of 0.981, with exceptional specificity at 0.997, while its sensitivity stands at 0.627. Interestingly, the RF maintains strong precision, delivering values of 0.956 and 0.959 with GA and PSO optimization, respectively.

The remarkable performance of the XGBoost+ANN hybrid model can be attributed to XGBoost's efficient handling of structured data and ANNs' ability to capture complex relationships in unstructured data. This combination leverages their complementary strengths. Meticulous parameter optimization, combined with strategies to address class imbalance, significantly enhances the predictive accuracy of the XGBoost+ANN hybrid model in bankruptcy prediction. Notably, the focused use of selected features allows XGBoost to concentrate on the most relevant information, reducing overfitting and achieving improved model accuracy, especially when fine-tuned through GA.

Conversely, the RF method stands out for its precision across different scenarios and optimization techniques (GA and PSO). Utilizing selected features boosts RF's potential to identify interactions and data noise, while optimization via GA and PSO fine-tunes RF's parameters to better suit the extensive feature set, resulting in superior precision. In summary, the XGBoost+ANN hybrid method exhibits superior performance in AUC, accuracy, sensitivity, and specificity, particularly when optimized with GA and PSO. Meanwhile, RF shows remarkable precision, especially with PSO optimization. In contrast, SVM displayed comparatively weaker performance.

The comparative performance of models using all features versus a selected subset of features is presented in Figures 4 to 7, illustrating AUC values through ROC curves.

In Figure 4, with GA optimization for all features, RF displays limited discriminative ability, indicated by an AUC of 0.66. Conversely, SVM shows poor performance

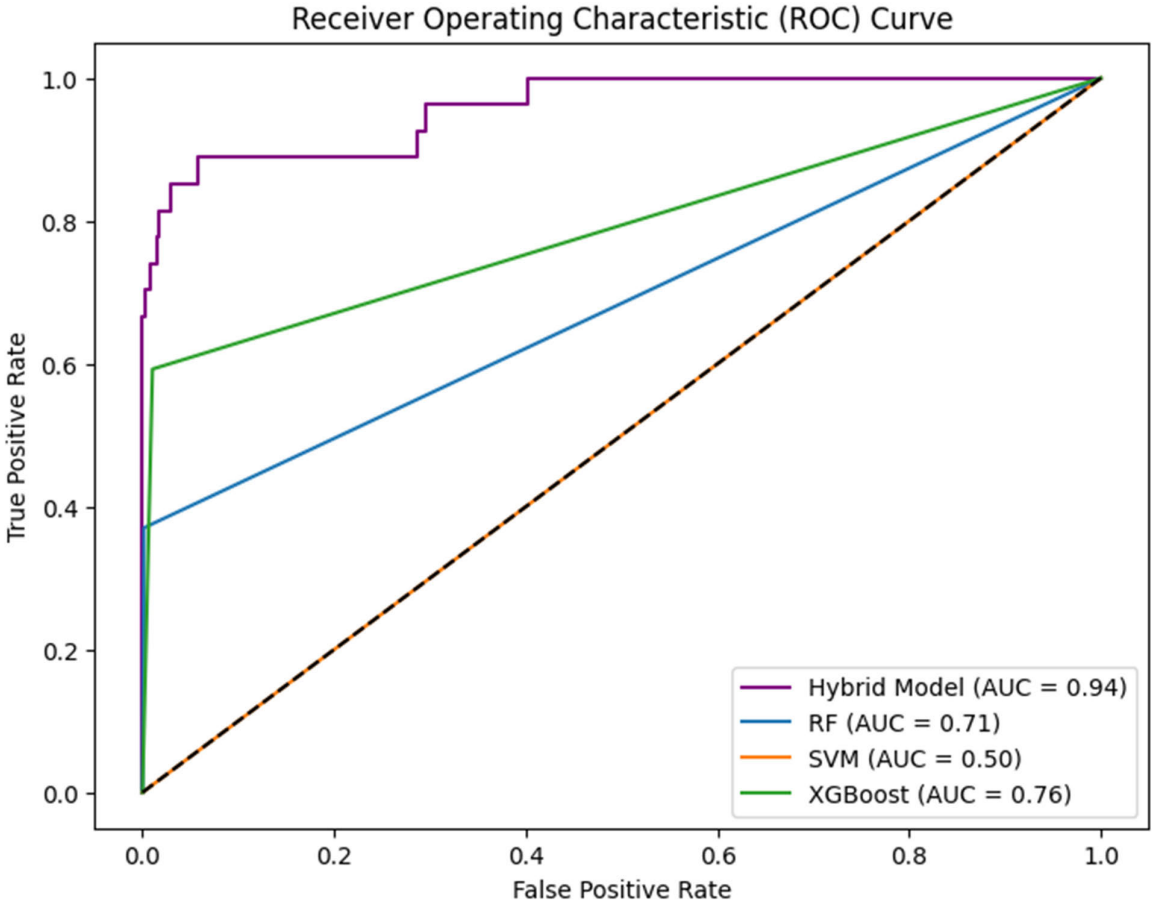


FIGURE 7. RF, SVM, XGBoost, and XGBoost+ANN with selected features and parameters optimized by PSO.

TABLE 7. Comparison of the proposed hybrid model XGBoost+ANN with earlier research using a Polish company dataset.

	Model	Acc	AUC	Feature Selection	Parameter Optimization	Data Balanced
Earlier Research Models	MOCS [9]	-	0.869	No	No	Yes
	ANN+PSO+MSE [24]	0.970	-	No	Yes	No
	ANN+CSO+MSE [24]	0.970	-	No	Yes	No
	CS-XGB [13]	0.979	0.955	No	No	No
	AP (70%) +LR [7]	-	0.735	No	No	No
	AP (60%) +SVM [7]	-	0.917	No	No	No
	K-means (6) +SVM [7]	-	0.917	No	No	No
	IFNA+backflow XGB [28]	0.975	0.919	No	Yes	Yes
Proposed Models	**XGBoost+ANN$_{(GA)}$**	**0.983**	**0.958**	No	Yes	No
	XGBoost+ANN$_{(PSO)}$	0.982	0.939	No	Yes	No
	XGBoost+ANN$_{(GA)}$	0.981	0.953	Yes	Yes	No
	XGBoost+ANN$_{(PSO)}$	0.981	0.936	Yes	Yes	No

with an AUC of 0.50, potentially due to sensitivity to noise or complex underlying patterns. XGBoost, with a moderate AUC of 0.77, demonstrates a reasonable ability to differentiate between classes. The XGBoost+ANN hybrid model, however, excels with an AUC of 0.96, showcasing very high discriminatory power and robust performance.

Figure 5, using PSO optimization for all features, shows an improvement in RF's performance to an AUC of 0.69, implying that PSO aids in more effective model parameter tuning. The SVM, with an AUC of 0.50, struggles to discern patterns. XGBoost shows reasonable performance with an AUC of 0.78, and the XGBoost+ANN hybrid model again excels with an AUC of 0.94, characterized by a notable improvement in the ROC curve, highlighting the effective synergy between boosting algorithms and neural networks.

Figures 6 and 7, featuring the use of GA and PSO with selected features, respectively, exhibit similar trends. The hybrid model consistently shows enhanced discriminative power across different optimization strategies, as evidenced by the marked improvements in AUC values. The analysis underscores the effectiveness of hybrid machine learning methods compared to single and ensemble methods in

handling the unbalanced Polish company dataset. Overall, optimization with GA tends to yield better results than with PSO. Interestingly, feature selection with GA had a minimal impact on performance, demonstrating the robustness of the hybrid models in various configurations.

D. PERFORMANCE COMPARISON

A comparative analysis presented in Table 7 between our proposed hybrid model and models from previous research, all applied to an unbalanced dataset, reveals that the XGBoost+ANN hybrid model outshines all others in terms of AUC (0.958) and accuracy (0.983).

This high performance was particularly notable when no feature selection was applied and parameters were optimized using Genetic Algorithms (GA). Notably, the CS-XGB model [13] achieved the second-highest AUC score of 0.955, despite not employing feature selection, parameter optimization, or data balancing techniques.

Conversely, the lowest AUC recorded was 0.735 for the AP (70%) + LR method [7], and the lowest accuracy was observed at 0.970 for the ANN+PSO+MSE and ANN+CSO+MSE [24] models. It's worth mentioning that models like MOCS [9] and IFNA+backflowXGB [28] were applied to balanced datasets, utilizing a strategic combination of over-sampling (SMOTE and ADASYN) and under-sampling (RUS and Tomek) methods. This approach to dataset balancing sets these models apart from the unbalanced dataset scenarios that our proposed hybrid model explored.

V. CONCLUSION

Predicting bankruptcy is crucial for the early warning and sustainability of companies. This research contributes significantly to this domain by introducing the hybrid method XGBoost+ANN, particularly focusing on enhancing prediction performance in unbalanced datasets. Our findings, substantiated by experiments using the Polish company dataset and comparisons with various benchmark models, have validated the effectiveness of these hybrid approaches.

A key highlight from our study is the exceptional performance of the XGBoost+ANN hybrid model. In comparative analysis, this model significantly outperformed others, achieving an AUC of 0.958 and an accuracy of 0.983. These results are even more impressive considering they were achieved without feature selection and with parameters optimized using GA.

The study's approach to class imbalance, employing GA and PSO for parameter optimization without resorting to complex resampling methods, proved effective. The XGBoost+ANN model, in particular, demonstrated a robust capability to handle unbalanced datasets, outshining traditional models in both accuracy and AUC metrics.

Despite these promising results, the research has areas for further exploration. The consistent use of ANN in hybrid models suggests the potential for investigating different combinations and approaches to refine accuracy further. Additionally, the effectiveness of GAs in feature selection could be enhanced with alternative methods. Future research will focus on these areas, including the application of advanced deep learning methods like LSTM and GRU and extending validations to diverse datasets like the ongoing Malaysian bankruptcy dataset.

REFERENCES

[1] S. Shetty, M. Musa, and X. Brédart, "Bankruptcy prediction using machine learning techniques," *J. Risk Financial Manage.*, vol. 15, no. 1, p. 35, Jan. 2022, doi: 10.3390/jrfm15010035.

[2] Y. Shi and X. Li, "An overview of bankruptcy prediction models for corporate firms: A systematic literature review," *Intangible Capital*, vol. 15, no. 2, p. 114, Oct. 2019, doi: 10.3926/ic.1354.

[3] T. M. Alam, K. Shaukat, M. Mushtaq, Y. Ali, M. Khushi, S. Luo, and A. Wahab, "Corporate bankruptcy prediction: An approach towards better corporate world," *Comput. J.*, vol. 64, no. 11, pp. 1731–1746, Nov. 2019, doi: 10.1093/comjnl/bxaa056.

[4] G. Perboli and E. Arabnezhad, "A machine learning-based DSS for mid and long-term company crisis prediction," *Expert Syst. Appl.*, vol. 174, Jul. 2021, Art. no. 114758, doi: 10.1016/j.eswa.2021.114758.

[5] D. Boughaci and A. A. K. Alkhawaldeh, "Appropriate machine learning techniques for credit scoring and bankruptcy prediction in banking and finance: A comparative study," *Risk Decis. Anal.*, vol. 8, nos. 1–2, pp. 15–24, May 2020, doi: 10.3233/RDA-180051.

[6] T. Le, "A comprehensive survey of imbalanced learning methods for bankruptcy prediction," *IET Commun.*, vol. 16, no. 5, pp. 433–441, Mar. 2022, doi: 10.1049/cmu2.12268.

[7] C. Tsai, "Two-stage hybrid learning techniques for bankruptcy prediction*," *Stat. Anal. Data Mining, ASA Data Sci. J.*, vol. 13, no. 6, pp. 565–572, Dec. 2020, doi: 10.1002/sam.11482.

[8] D. Veganzones and E. Séverin, "An investigation of bankruptcy prediction in imbalanced datasets," *Decis. Support Syst.*, vol. 112, pp. 111–124, Aug. 2018, doi: 10.1016/j.dss.2018.06.011.

[9] Y. Zelenkov and N. Volodarskiy, "Bankruptcy prediction on the base of the unbalanced data using multi-objective selection of classifiers," *Expert Syst. Appl.*, vol. 185, Dec. 2021, Art. no. 115559, doi: 10.1016/j.eswa.2021.115559.

[10] H. Son, C. Hyun, D. Phan, and H. J. Hwang, "Data analytic approach for bankruptcy prediction," *Expert Syst. Appl.*, vol. 138, Dec. 2019, Art. no. 112816, doi: 10.1016/j.eswa.2019.07.033.

[11] S. S. Devi and Y. Radhika, "A survey on machine learning and statistical techniques in bankruptcy prediction," *Int. J. Mach. Learn. Comput.*, vol. 8, no. 2, pp. 133–139, Apr. 2018, doi: 10.18178/ijmlc.2018.8.2.676.

[12] K. UlagaPriya and S. Pushpa, "A comprehensive study on ensemble-based imbalanced data classification methods for bankruptcy data," in *Proc. 6th Int. Conf. Inventive Comput. Technol. (ICICT)*, Jan. 2021, pp. 800–804, doi: 10.1109/ICICT50816.2021.9358744.

[13] W. Yotsawat, K. Phodong, T. Promrat, and P. Wattuya, "Bankruptcy prediction model using cost-sensitive extreme gradient boosting in the context of imbalanced datasets," *Int. J. Electr. Comput. Eng. (IJECE)*, vol. 13, no. 4, p. 4683, Aug. 2023, doi: 10.11591/ijece.v13i4.pp4683-4691.

[14] T.-K. Chen, H.-H. Liao, G.-D. Chen, W.-H. Kang, and Y.-C. Lin, "Bankruptcy prediction using machine learning models with the text-based communicative value of annual reports," *Expert Syst. Appl.*, vol. 233, Dec. 2023, Art. no. 120714, doi: 10.1016/j.eswa.2023.120714.

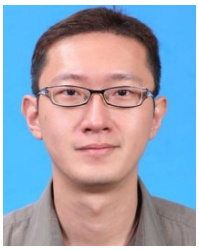
[15] A. Narvekar and D. Guha, "Bankruptcy prediction using machine learning and an application to the case of the COVID-19 recession," *Data Sci. Finance Econ.*, vol. 1, no. 2, pp. 180–195, 2021, doi: 10.3934/DSFE.2021010.

[16] T. Le, L. Hoang Son, M. Vo, M. Lee, and S. Baik, "A cluster-based boosting algorithm for bankruptcy prediction in a highly imbalanced dataset," *Symmetry*, vol. 10, no. 7, p. 250, Jul. 2018, doi: 10.3390/sym10070250.

[17] R. F. Brenes, A. Johannssen, and N. Chukhrova, "An intelligent bankruptcy prediction model using a multilayer perceptron," *Intell. Syst. with Appl.*, vol. 16, Nov. 2022, Art. no. 200136, doi: 10.1016/j.iswa.2022.200136.

[18] T. Le, M. T. Vo, B. Vo, M. Y. Lee, and S. W. Baik, "A hybrid approach using oversampling technique and cost-sensitive learning for bankruptcy prediction," *Complexity*, vol. 2019, pp. 1–12, Aug. 2019, doi: 10.1155/2019/8460934.
[19] T. Le, M. Lee, J. Park, and S. Baik, "Oversampling techniques for bankruptcy prediction: Novel features from a transaction dataset," *Symmetry*, vol. 10, no. 4, p. 79, Mar. 2018, doi: 10.3390/sym10040079.
[20] M. Zięba, S. K. Tomczak, and J. M. Tomczak, "Ensemble boosted trees with synthetic features generation in application to bankruptcy prediction," *Expert Syst. Appl.*, vol. 58, pp. 93–101, Oct. 2016, doi: 10.1016/j.eswa.2016.04.001.
[21] A. H. Uddin, Y.-L. Chen, B. Borkatullah, M. S. Khatun, J. Ferdous, P. Mahmud, J. Yang, C. S. Ku, and L. Y. Por, "Deep-learning-based classification of Bangladeshi medicinal plants using neural ensemble models," *Mathematics*, vol. 11, no. 16, p. 3504, Aug. 2023, doi: 10.3390/math11163504.
[22] R. Kanapickienė, T. Kanapickas, and A. Nečiūnas, "Bankruptcy prediction for micro and small enterprises using financial, non-financial, business sector and macroeconomic variables: The case of the Lithuanian construction sector," *Risks*, vol. 11, no. 5, p. 97, May 2023, doi: 10.3390/risks11050097.
[23] L. Yang and A. Shami, "On hyperparameter optimization of machine learning algorithms: Theory and practice," *Neurocomputing*, vol. 415, pp. 295–316, Nov. 2020, doi: 10.1016/j.neucom.2020.07.061.
[24] S. A.-D. Safi, P. A. Castillo, and H. Faris, "Cost-sensitive metaheuristic optimization-based neural network with ensemble learning for financial distress prediction," *Appl. Sci.*, vol. 12, no. 14, p. 6918, Jul. 2022, doi: 10.3390/app12146918.
[25] S. Aly, M. Alfonse, M. I. Roushdy, and A. B. M. Salem, "Developing an intelligent system for predicting bankruptcy," *J. Theory Appl. Inf. Technol.*, vol. 100, no. 7, pp. 2068–2087, Apr. 2022.
[26] M. Brygała, "Consumer bankruptcy prediction using balanced and imbalanced data," *Risks*, vol. 10, no. 2, p. 24, Jan. 2022, doi: 10.3390/risks10020024.
[27] S. H. Syed Nor, S. Ismail, and B. W. Yap, "Personal bankruptcy prediction using decision tree model," *J. Econ., Finance Administ. Sci.*, vol. 24, no. 47, pp. 157–170, Jun. 2019, doi: 10.1108/JEFAS-08-2018-0076.
[28] S. Wei, D. Yang, W. Zhang, and S. Zhang, "A novel noise-adapted two-layer ensemble model for credit scoring based on backflow learning," *IEEE Access*, vol. 7, pp. 99217–99230, 2019, doi: 10.1109/ACCESS.2019.2930332.
[29] S. U. Haq, S. U. Bazai, A. Fatima, S. Marjan, J. Yang, L. Y. Por, M. Anjum, S. Shahab, and C. S. Ku, "Reseek-arrhythmia: Empirical evaluation of ResNet architecture for detection of arrhythmia," *Diagnostics*, vol. 13, no. 18, p. 2867, Sep. 2023, doi: 10.3390/diagnostics13182867.
[30] H. Liang, M. Zhang, and H. Wang, "A neural network model for wildfire scale prediction using meteorological factors," *IEEE Access*, vol. 7, pp. 176746–176755, 2019, doi: 10.1109/ACCESS.2019.2957837.
[31] F. Itano, M. A. de Abreu de Sousa, and E. Del-Moral-Hernandez, "Extending MLP ANN hyper-parameters optimization by using genetic algorithm," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2018, pp. 1–8, doi: 10.1109/IJCNN.2018.8489520.
[32] S. Hosseini and B. M. H. Zade, "New hybrid method for attack detection using combination of evolutionary algorithms, SVM, and ANN," *Comput. Netw.*, vol. 173, May 2020, Art. no. 107168, doi: 10.1016/j.comnet.2020.107168.
[33] R. Sikora and S. Piramuthu, "Framework for efficient feature selection in genetic algorithm based data mining," *Eur. J. Oper. Res.*, vol. 180, no. 2, pp. 723–737, Jul. 2007, doi: 10.1016/j.ejor.2006.02.040.
[34] M. A. Albadr, S. Tiun, M. Ayob, and F. Al-Dhief, "Genetic algorithm based on natural selection theory for optimization problems," *Symmetry*, vol. 12, no. 11, p. 1758, Oct. 2020, doi: 10.3390/sym12111758.
[35] F. Barboza, H. Kimura, and E. Altman, "Machine learning models and bankruptcy prediction," *Expert Syst. Appl.*, vol. 83, pp. 405–417, Oct. 2017, doi: 10.1016/j.eswa.2017.04.006.
[36] V. F. Rodriguez-Galiano, B. Ghimire, J. Rogan, M. Chica-Olmo, and J. P. Rigol-Sanchez, "An assessment of the effectiveness of a random forest classifier for land-cover classification," *ISPRS J. Photogramm. Remote Sens.*, vol. 67, pp. 93–104, Jan. 2012, doi: 10.1016/j.isprsjprs.2011.11.002.
[37] X. Li and Y. Sun, "Stock intelligent investment strategy based on support vector machine parameter optimization algorithm," *Neural Comput. Appl.*, vol. 32, no. 6, pp. 1765–1775, Mar. 2020, doi: 10.1007/s00521-019-04566-2.
[38] J. Huang and C. X. Ling, "Using AUC and accuracy in evaluating learning algorithms," *IEEE Trans. Knowl. Data Eng.*, vol. 17, no. 3, pp. 299–310, Mar. 2005, doi: 10.1109/TKDE.2005.50.

UMMEY HANY AINAN received the B.Sc. degree from the International Islamic University of Chittagong, Bangladesh. She is currently pursuing the master's degree with the Universiti Malaya. Her research interests encompass various domains, including bank data processing, machine learning (such as gradient boosting and support vector machines), artificial intelligence (such as artificial neural networks), and genetic algorithms.



LIP YEE POR (Senior Member, IEEE) received the B.Sc., M.Sc., and Ph.D. degrees from Universiti Malaya, Malaysia. He currently holds the position of an Associate Professor with the Faculty of Computer Science and Information Technology, Universiti Malaya. His research interests encompass various aspects of information security and quality assurance (NEC 2020: 0611), including authentication, graphic passwords, PIN-entry, cryptography, data hiding, steganography, and watermarking. Additionally, he specializes in machine learning (NEC 2020: 0613), with expertise in extreme learning machines, support vector machines, deep learning, long-short-term memory, computer vision, and AIoT.

YEN-LIN CHEN (Senior Member, IEEE) received the B.S. and Ph.D. degrees in electrical and control engineering from the National Chiao Tung University, Hsinchu, Taiwan, in 2000 and 2006, respectively. From February 2007 to July 2009, he was an Assistant Professor with the Department of Computer Science and Information Engineering, Asia University, Taichung, Taiwan. From August 2009 to January 2012, he was an Assistant Professor with the Department of Computer Science and Information Engineering, National Taipei University of Technology, Taipei, Taiwan, where he was an Associate Professor, from February 2012 to July 2015. Since August 2015, he has been a Full Professor with the National Taipei University of Technology. His research interests include artificial intelligence, intelligent image analytics, embedded systems, pattern recognition, intelligent vehicles, and intelligent transportation systems. His research results have published on over 100 journals and conference papers. He is a fellow of IET; and a member of ACM, IAPR, and IEICE.

JING YANG (Graduate Student Member, IEEE) received the Bachelor of Engineering degree majoring in navigation technology from Shandong Jiaotong University, in 2022. He is currently pursuing the master's degree in data science with Universiti Malaya. His primary research interests lie in the fields of medical image processing and deep learning.

CHIN SOON KU received the Ph.D. degree from Universiti Malaya, Malaysia, in 2019. He is currently an Assistant Professor with the Department of Computer Science, Universiti Tunku Abdul Rahman, Malaysia. His research interests include AI techniques (such as genetic algorithms), computer vision, decision support tools, graphical authentication (authentication, picture-based passwords, and graphical passwords), machine learning, deep learning, speech processing, natural language processing, and unmanned logistics fleets.