# Early Detection of Stroke for Ensuring Health and Well-Being Based on Categorical Gradient Boosting Machine

**Isaac Kofi Nti[1,2*], Owusu Nyarko-Boateng[1], Justice Aning[3], Godfred Kusi Fosu[3], Henrietta Adjei Pokuaa[3] & Frimpong Kyeremeh[4]**

[1]Department of Computer Science and Informatics,
University of Energy and Natural Resources, Post Office Box 214, Sunyani, Ghana
[2]School of Information Technology, University of Cincinnati,
2610 University Circle, Cincinnati, Ohio 45221USA
[3]Department of Computer Science, Sunyani Technical University,
Syi-Ksi Highway, Sunyani, Ghana
[4]Department of Electrical and Electronic Engineering,
Sunyani Technical University, Syi-Ksi Highway, Sunyani, Ghana
*E-mail: ntious1@gmail.com

**Abstract.** Stroke is believed to be among the leading causes of adult disability worldwide. It is wreaking havoc on African people, families, and governments, with ramifications for the continent's socio-economic development. On the other hand, stroke research output is insufficient, resulting in a dearth of evidence-based and context-driven guidelines and strategies to combat the region's expanding stroke burden. Indeed, for African and other developing economies to meet the UN Sustainable Development Goals (SDGs), particularly SDG 3, which aims to guarantee healthy lifestyles and promote well-being for people of all ages, the issue of stroke must be addressed to reduce early death from non-communicable illnesses. This study sought to create a robust predictive model for early stroke diagnosis using an understandable machine learning (ML) technique. We implemented a categorical gradient boosting machine model for early stroke prediction to protect patients' health and well-being. We compared the effectiveness of our proposed model to existing state-of-the-art machine learning models and previous studies by empirically testing it on a real-world public stroke dataset. The proposed model outperformed the others when compared to the other methods using the research data, achieving the maximum accuracy (96.56%), the area under the curve (AUC) (99.73%), F1-measure (96.68%), recall (99.24%), and precision (93.57%). Functional outcome prediction models based on machine learning for stroke were verified and shown to be adaptable and helpful.

**Keywords**: *CatBoost; gradient boosting machine; health; stroke; stroke prediction; well-being.*

## 1 Introduction

Stroke is a neurological condition caused by ischemia or bleeding of the brain arteries, usually resulting in various physical and cognitive problems that make it

Received February 3rd, 2022, Revised September 8th, 2022, Accepted for publication October 13th, 2022
 ISSN: 2337-5787, DOI: 10.5614/itbj.ict.res.appl.2022.16.3.8

challenging to function [1-3]. It is the third leading cause of death and severe long-term disability in most advanced and emerging countries worldwide [4,5]. In addition, stroke is one of the most common causes of adult impairment, resulting in loss of capacity. According to statistics from the World Health Organization (WHO), 15 million people died in 2015 from cardiac disease, with 6.24 million deaths from stroke [6]. Following heart disease, stroke was the second most common cause of death and one of the top-five causes of disability in 2020.

Anyone can have a stroke anywhere at any time [7,8]. About a hundred years ago, stroke was relatively uncommon in Africa. However, of late, it is prevalent in most parts of Africa, with a yearly incidence rate of 316 per 100,000, a yearly prevalence rate of 1,460 per 100,000, and a three-year death toll of up to 84% [9-11].

Furthermore, many Africans get it at the pinnacle of their career and societal contribution. According to several studies [4,12-14], increased incidence of cardio-metabolic risk features, populace expansion, decreased corporal activity, and other lifestyle deviations contribute to the rising stroke burden in Africa. Stroke is a leading public health problem with substantial economic and social ramifications. In Ghana, it is one of the top-three causes of death and a prominent cause of adult medical hospitalization [4,12,15]. Stroke is a significant public health concern in Ghana, but no contemporary research has been done on the disease using machine learning for early detection. Early detection of several strokes causes could help prevent the escalating numbers of stroke incidents in Ghana and other West African countries.

Predicting a person's likelihood of having a stroke is critical for early intervention and treatment. With the advancement of artificial intelligence, technologies such as machine learning (ML) can assist patients and clinicians in detecting stroke symptoms in an early stage [16]. ML application in different health sectors has provided accurate and speedy prediction results and has hence become a powerful tool of late in health settings. Its application in stroke detection allows stroke patients to receive tailored clinical therapy in most developed countries [2,3,17]. However, although ML in health care is rising, some developing countries like Ghana are yet to receive enough scientific attention, despite a clear need for research. Hence, this study proposes a method for predicting early stroke disorders by analyzing blood pressure, body mass index (BMI), heart disease, gender, and other factors, such as smoking and age. Specifically, our goal was to create a novel prediction model for early stroke identification using an understandable ML algorithm and to assess the prediction accuracy and relevance of ML models. In summary, the following are our significant contributions:

1. Propose a robust feature engineering scheme to discover new (unknown) possible stroke risk factors.
2. Adapt the ordered target statistics technique in the categorical gradient boosting algorithm (CatBoost) to convert categorical information to numerical features during training to reduce data processing complexity.
3. Develop an integrated ML method for stroke prediction that significantly outperforms existing state-of-the-art algorithm and studies.
4. This study is one of the few from developing countries like Ghana to propose a machine learning-based smart stroke prediction model to assist stakeholders in making informed decisions.

We believe this study's outcome will help individuals with no or insufficient knowledge of stroke to detect the early likelihood of the disease to seek early medical help. Also, it will help facilitate the work of the limited number of clinicians in the country (Ghana). The remaining sections of the study are categorized as follows. First, we present the literature survey in Section 2. Next, Section 3 discusses the study methodology, while Section 4 presents the results and discussions. Finally, Section 5 contains the study conclusions and future works.

## 2 Literature Survey

### 2.1 Machine Learning

Studying algorithms that can learn quickly and improve without explicitly programming is known as the field of machine learning. ML techniques can be classified as supervised, unsupervised, or deep learning. Supervised learning develops a model that predicts the outcome by mapping one or more independent variables to a target (output) based on observations. It is usually grouped into two: classification and regression. The term 'classification' refers to using predictors to categorize a discrete target variable. The relationship between the target variable and the predictors is investigated using regression analysis [2,18,19]. Unsupervised learning is used to cluster observations to create clusters based on likeness. Unsupervised learning techniques include clustering, association analysis, and dimensionality reduction. Finally, deep learning creates a computational model with multiple processing layers that learns a data set incrementally from raw data [2].

### 2.2 Related Works

A hybrid machine learning model was proposed for stroke prediction based on an incomplete and unbalanced medical dataset [20]. Random forest (RF) was adopted for missing data replacement and hyperparameter optimization

(AutoHPO), based on a deep neural network (DNN), was used to predict stroke. Using 43,400 biological data records achieved 71.6% accuracy and 67.4% sensitivity. Likewise, Lau, *et al*. [21] looked into automatic movement recognition using the SVM ML algorithm and data from kinematic sensors. Seven patients walked in various environments, including stair_descent, stair_ascent, level_ground, downslope, and upslope. The support vector machine (SVM) outperformed artificial neural networks (ANN) and radial basis function neural networks, improving accuracy \ to 97.5% (RBF). Finally, SVM was used to predict stroke with Cardiac Health Study (CHS) data. The study proposed a novel automatic feature selection method that selects robust features based on the conservative mean [22]. However, this resulted in many vectors that degraded the model's performance. Likewise, different (decision trees (DT), ANN, SVM and principal component analysis (PCA)) ML techniques were used to detect stroke in the CHS dataset [23]. Nevertheless, their dataset had fewer features, limiting its ability to examine the significant factors that predict stroke.

Another study [24] has proposed a stroke prediction algorithm based on an improvised RF algorithm. According to the authors, this method outperformed existing algorithms. However, the study was limited to only a few types of strokes and does not apply to any new stroke types in the future. DT, RF, and multi-layer perceptron (MLP) were used for stroke prediction. The accuracy among them was quite similar, with only minor differences (DT = 74.31%, RF = 74.53%, and MLP = 75.02%). Based on the results, MLP appeared to be more accurate than the other two methods [25]. The accuracy score was the only metric used to calculate the performance; however, studies [16,26,27] have shown that it may not always give favorable results in ML applications. Three ML algorithms (DT, MLP and Jrip) were hybrids to predict the likelihood of having a stroke and were tested on a dataset gathered from Saudi Arabia's Ministry of National Guards Health Affairs Hospitals [28]. The model attained an accuracy of 95%, however, with a high computational time compared to other works in the same field. Naive Bayes, Neural Networks, and DT models were developed in [29]. DT outperformed the other two algorithms by 75%. Nevertheless, their model is not practicable in real-world scenarios based on their evaluation metric (confusion matrix). Jamthikar, *et al.* [30] compared ML algorithms and statistical models for accurate stroke prediction [30]. The study compared thirteen different statistically determined risk calculators; the obtained experimental findings showed that the ML-based model (SVM) showed superior prediction ability. However, the study dataset was ethnically biased, which may hinder the generalization of the model according to the authors.

Similarly, an ANN trained with the Back-propagation technique was proposed for detecting thrombo-embolic stroke illness [31]. It achieved a prediction accuracy of 89%; however, the computational time was complex due to the

complicated structure and increased number of neurons. Likewise, integrated ML algorithms (SVM, LightGBM, MLP and logistic regression) were applied to predict neurological deterioration in atrial fibrillation (AF)-related stroke cases [32]. LightGBM exhibited the best performance among the ML models tested, with an AUROC score of 77.2%. In another study, RF, extreme gradient boosting models and generalized logistic regression (GLR) were applied to predict patients with acute ischemic stroke [33]. GLR outperformed all other ML algorithms with an AUC of 0.86. Also, the report showed that the top-two critical factors for predicting stroke were the health stroke scale at admission and age.

It is evident from the above studies that machine learning (ML) algorithms have become more prevalent in the medical industry, owing to their ability to include and evaluate a large number of factors and large amounts of data. In contrast to conventional statistical models, machine learning models can forecast complicated clinical outcomes influenced by various factors. However, as seen above, ML application in the healthcare industry is minimal in developing countries compared with developed countries. Therefore, machine learning models must be applied in developing nations to help reduce mortality due to stroke. Hence, our research aimed to create an interpretable machine learning model capable of predicting early stroke symptoms from real-world data.

## 3 Methodology

Figure 1 shows the data-flow diagram of our stroke prediction model. It consists of three (3) phases: (i) data preprocessing – data cleaning, feature selection and data partitioning, (ii) building of learning algorithms for prediction, and (iii) model evaluation. We explain each step in the following sub-section.

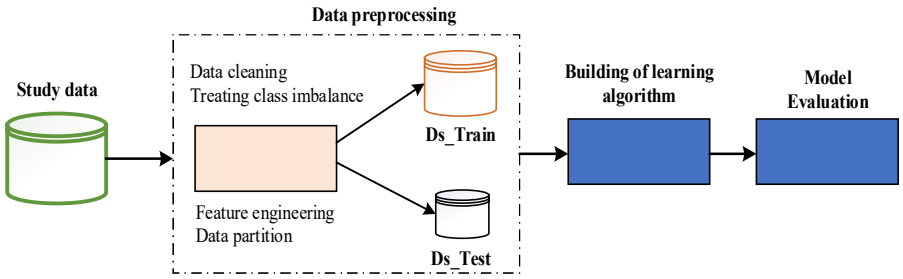


**Figure 1** Study data-flow diagram.

### 3.1 Dataset Preprocessing

The dataset used in this study was downloaded from Kaggle.[1] The features and description of the downloaded dataset are shown in Table 1. We first preprocessed our data by treating missing values. Clinical data frequently includes significant omissions because patients opting out of surveys, data gathering problems, and other factors.

**Table 1** Features of study dataset used in all experiments.

| Name | Description | Type |
|---|---|---|
| age | Age of the patient | Numerical |
| Avg-glucose-level | Blood glucose level average | |
| BMI | Body mass index | |
| Hypertension * | If the patient does not have hypertension, the answer is 0. 1 If the patient has. | Categorical |
| gender | Patient's gender | |
| Heart-disease * | If the patient has no cardiac problems, the answer is 0. However, if the patient has a cardiac condition, then 1. | |
| Residence_type * | The patient's residence, whether city or rural area | |
| Ever_married ** | Marital status of the patent | |
| Work-type ** | The kind of work the patient does | |
| Smoking-status | The patient's smoking status | |
| stroke | If the patient experienced a stroke, the answer is 1; otherwise, the answer is 0. | |

A single asterisk (*) denotes that one-hot encoding is utilized in CatBoost, RF and SVM, whereas a double asterisk (**) denotes that the functionality is only available in CatBoost.

In treating missing values, we replaced each value with the mean of the observed features in the column. Imbalanced data can significantly reduce the accuracy of the ML model; hence, the random over-sampling technique was implemented using the imbalanced-learn Python module to balance the study dataset.

Next was feature selection, because building an appropriate clinical data model requires the selection of key elements. Only a tiny group of characteristics is significantly related to stroke prediction. Therefore, the study adopted a regularized random forest (RRF) algorithm for feature selection. The typical method for predicting stroke is manually selecting characteristics based on risk factors from medical and clinical research. Finally, the clean dataset had a size of (5110, 11); we partitioned it into 80% training data and 20% testing data.

[1] https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset

## 3.2 Learning Algorithms For Prediction

### 3.2.1 CatBoost (Categorical Boosting) Model

Gradient boosting is a strong ML technique that produces cutting-edge outcomes in various real-world applications. It is a fundamental technique for learning tasks with mixed features, noisy data, and complicated dependencies: web search, weather forecasting, recommendation schemes, and various other applications [34,35]. It is supported by theoretical findings showing how robust predictors can be constructed by iteratively merging less effective models (base predictors) using a greedy technique corresponding to gradient descent in a function space. This study adopted a novel gradient boosting technique (CatBoost) that handles categorical features very well and takes advantage of dealing with them through training instead of preprocessing [34,36].

When choosing the tree structure, it also employs an innovative schema for leaf value estimation which aids in reducing overfitting. CatBoost is a categorical data gradient boosting library. It does not employ binary substitution for categorical values; rather, it transposes the dataset and estimates the average label-value for every case with a similar category value [35,37,38]. The gradient boosting approach employs oblivious decision trees as base predictors to create CatBoost. The DTs are used for classification, with each tree representing a feature space split and an output value. Each tree level is split according to a decision rule (DR). Each DR considers a pair *r = (i, j)* consisting of a feature index *i = 1,...,n* and a threshold value $j \in \Re$. A collection of feature vectors X can be divided into two distinct subgroups, $X^A$ and $X^B$, using this decision rule so that for each, $x = (x^1, \cdots, x^n) \in X$, we have Eq. (1).

$$x \in \begin{cases} X^A, & if\ x^i \le j \\ X^B, & if\ x^i > j \end{cases} \tag{1}$$

The decision rule is applied to *s* disjoint sets $X_{(1)}, \cdots, X_{(s)} \in \Re^n$ to get *2s* disjoint sets $X^A_{(1)}, X^B_{(1)}, \cdots, X^A_{(s)}, X^B_{(s)}$. Assuming a collection of sets $\left(N = \{X_{(1)}, \cdots, X_{(s)}\}\right)$ and a target function $(f: \Re^n \to \Re)$, the decision rule is expressed as Eq. (2).

$$\text{argmin}_a\{P(a, f, N)\} \tag{2}$$

where *N* is a function that evaluates the decision rule and the collection *M* in terms of their optimality concerning the goal function *f*. In oblivious DT *(P)* is expressed by Eq. (3).

$$P(a, f, N) = \left\{\left(\frac{1}{\sum_{k=1}^{s} |X_{(k)}|}\right)\left(\sum_{k=1}^{s} |X^A_{(k)}| Var\left(f\left(X^A_{(k)}\right)\right) + |X^B_{(k)}| Var\left(f\left(X^B_{(k)}\right)\right)\right)\right\} \tag{3}$$

where $f(X_{(k)})$ is the target scores corresponding to sample $X_{(k)}$. Prokhorenkova, *et al.* [37] established a novel strategy called ordered boosting to avoid gradient bias through theoretical study. Algorithm 1 is the pseudo-code for ordered boosting.

**Algorithm 1**

$Input$: $\{(X_k, Y_k)\}_{k=1}^{v}$ ordered according to $\sigma$, total trees I;
$\sigma \leftarrow random\ permutation\ of\ [1, v]$
$\mathrm{U}_i \leftarrow 0\ for\ i\ = 1, \cdots, v$
$for\ t \leftarrow 1\ to\ I\ do$
$for\ i \leftarrow 1\ to$ v $do$
$r_i \leftarrow y_i - U_{\sigma(i)-1}(X_i)$;
$for\ i\ \leftarrow 1\ to$ v $do$
$\Delta U \leftarrow LearnModel[(X_i, r_j): \sigma(j) \leq i]$
$\mathrm{U}_i \leftarrow U_i + \Delta U$
$return\ \mathrm{U}_v$

It is worth noting that *Ui* was trained without using the *Xi* example, and each *Ui* had an identical tree topology.

Our training dataset (DS_Train) was permuted *s* times in CatBoost. We selected a random permutation and calculated its gradients to improve the algorithm's resilience. The same permutations were used to compute categorical feature statistics. We trained various models with different permutations, so we did not worry about overfitting. As illustrated above, we trained *n* different models ($M_i$) for each permutation (Algorithm 1). This implies that for each permutation, we had to store and recalculate $O(n_2)$ approximations: for each model $M_i$, we updated $M_i(X_1)$..., $M_i(X_i)$.

As a result, this operation had an O complexity ($sn^2$). Instead of storing and updating $O(n^2)$ values $M_i(X_i)$, we kept values in our practical implementation, which decreased the complexity of one tree building to $o(sn)$; we kept values $M_i^{'}(X_j), i = 1, \cdots, [log_2(n)], j < 2^{i+1}$, where $M_i^{'}(X_j)$ is the rough calculation for the same *j* based on the first $2^i$ samples. Then, the number of predictions $M_i^{'}(X_j)$ was not larger than $\sum_{0 \leq i \leq l \quad _2(n)} 2^{i+1} < 4n$.

The gradient on the example $X_k$ on the basis of the approximation used for picking a tree structure was estimated $M_i^{'}(X_k)$, where $i = [log_2(k)]$. The hyperparameters of the CatBoost for this study were: iterations = 210, feature_border_type = GreedyLogSum, bayesian_matrix_reg = 0.11, l2_leaf_reg = 6, learning_rate = 0.001, score_function = Cosine, leaf_estimation_iterations = 12 and max_leaves = 2.

### 3.2.2 Comparison Models

Based on Section 2.2, we adopted the two most used machine learning algorithms for stroke prediction compared to our proposed CatBoost model.

#### 3.2.2.1 Random Forests (RF)

RF combines several DT classifiers into a single, powerful model using Breiman's 'bagging' concept [39]. It employs a self-help approach to produce fresh training subsets from the original *K* training samples by periodically choosing random $v(v < K)$ sets of samples. Some samples may be gathered more than once throughout the overall selection procedure. Out-of-bag (OOB) data refers to the 36.8% of training data not sampled in each cycle of bagging random sampling. These uncollected data are not included in the model fitting during training, but they can be used to test the model's capacity to generalize.

The training sample is utilized to form random forests by building *v* buffeting decision or regression trees (CART). The test sample is then classified using a majority vote decision or the average return value. RF, generally, may attain strong generalization skills and low modification resistance without extra pruning because randomization can efficiently reduce model variance. For this study the RF hyperparameters were: criterion = entropy, max depth = 4, max features = log2, min_impurity_decrease = 0.0002 and n_estimators = 130. This study implemented the RF model using the Scikit learn library.

#### 3.2.2.2 Support Vector Machine (SVM)

The SVM ML technique was created by AT&T Bell Labs to reliably categorize binary features using a single hyperplane (H) [40]. It converts input characteristics (*v*) into a higher-dimensional space with the finest separating hyperplane. The hyperplane is a boundary between two classes constructed by maximizing the boundary between the support vectors (SV) of both classes [41,42].

In most cases, the hyperplane has various values for different feature dimensions, such as a point for one dimension and a straight line for two dimensions. H would be a plane if the dimensions were bigger than two. The best H has the most significant difference between classes. The largest width parallel to the H of the slice with no interior data points is called the margin. SV are the data points that are close to the dividing H. This study implemented the SVM model using the Scikit learn library.

### 3.3 Evaluation Metrics

Machine learning model performance may be measured using a variety of statistical approaches. However, for the sake of this research, we adopted six well-known metrics (see Table 2).

**Table 2** Adopted evaluation metrics.

| # | Technique | Formula | Description |
|---|---|---|---|
| 1 | Accuracy (ACC) | $ACC = \left\{ \frac{(TP+TN)}{(TP+FP+TN+FN)} \right\}$ | Expresses the proportion of correct forecasts to total predictions. TP = true positive, TN = true negative, FN = false negative, and FP = false positive. |
| 2 | Precision | $\text{Precision} = \left\{ \frac{TP}{TP+FP} \right\}$ | Assesses a model's ability to classify the positive label. A score closer to one (1) is preferable. |
| 3 | recall/ Sensitivity | $\text{Recall} = \left\{ \frac{TP}{TP+FN} \right\}$ | Describes how well a classification model classifies the positive class when the outcome is positive. A recall value of one (1) or less is preferable. |
| 4 | F-Score (FS) | $FS = \left\{ \frac{2 \times P \times R}{P+R} \right\}$ | Depicts the equilibrium between (P) and (R), i.e., P = consonant mean and R = sensitivity. |
| 5 | ROC | The receiver operator characteristic (ROC) is a probability curve that shows the TPR against the FPR at several threshold points, effectively separating the wanted signal from the unwanted. | |
| 6 | AUC | The area under the curve (AUC) summarizes the receiver operator characteristic (ROC) curve that assesses a model's capacity to differentiate between labels. AUC values of 1 indicate that the classifier is 'best', with 0.5 indicating 'random guessing'. | |

## 4 Experimental Results and Discussions

We conducted experiments to assess the performance of our proposed stroke prediction model. The experiments were run on an HP laptop (Spectre x360) with an Intel Core i7 CPU with 16.0 GB RAM. The Scikit[2] learn library and Python programming language were used in this study to implement all machine learning models. We present the results in the following sections. First, Figure 2 shows a plot of the study data: Figure 2(A) shows the unbalanced data, and Figure 2(B) shows the data after balancing with the random over-sampling technique.

[2] https://scikit-learn.org/

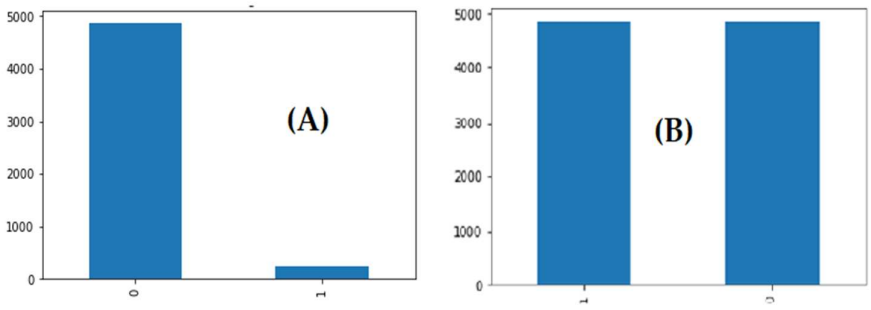


**Figure 2** Study dataset: unbalanced dataset (A), balanced dataset using a random over-sampling technique (B).

## 4.1 Feature Selection and Feature Importance Analysis

The top-*n* significant predictors contributing to the efficiency of the proposed model were selected using recursive feature elimination. This technique automatically detects possible risk factors without requiring lengthy medical studies to comprehend each one thoroughly. It enables a rapid approach to describing a novel illness and identifying its predictors before further research confirms them. In addition, this approach may potentially be utilized to identify previously unknown risk factors. We discovered the top features in our tests by rating the average of the conservative mean vectors in descending order over several random trials.

Figure 3 shows a plot of the top attributes of the study data. It is worth noting that the top properties recognized by medical research and those chosen by our feature selection algorithm have a lot in common. The outcome reveals that our system is capable of correctly recognizing risk variables. As a result, highly ranked traits that have yet to be clinically validated may be potential risk factors. For example, age and glucose level were the most significant features in predicting stroke, which agrees with previous literature [22,33,43].

Average glucose was a high-ranking component, although BMI is a well-known stroke risk factor. Furthermore, hypertension may be a significant risk factor since it indicates that a person's cerebrovascular activity is linked to stroke risk. Our findings also point to some other possible stroke risk factors, such as marital status and gender. More research on these characteristics may lead to better stroke prediction.

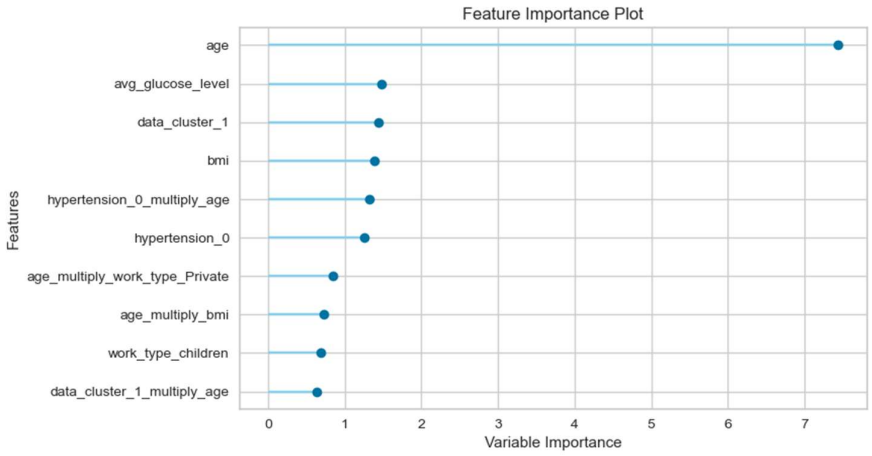


**Figure 3** Feature importance ranking.

## 4.2 Model Performance

The ten-fold CV method was conducted using the datasets discussed in Section 3.1 to test the success of our proposed model for predicting stroke. In addition, the proposed CatBoost model, RF and SVM methods were trained using a grid search hyperparameter optimization technique to optimize the parameters.

Table 3 shows the performance of our proposed model using ten-fold CV training. The cross-validation approach produces a reliable performance comparison because it uses data set components individually (i.e., folds) for the testing and training processes. Our technique attained an average AUC value of 99.73%, indicating that it can differentiate between a patient who is likely to get a stroke or not. Furthermore, our proposed method had an average accuracy of 96.56% in predicting stroke patients.

In addition, we observed an average recall score of 99.24%, indicating that our method can successfully detect all stroke cases with a low false-negative rate. Again, the proposed technique achieved average precision scores of 93.57%, which is the ratio of properly diagnosed stroke patients to classified stroke cases. Also, our model observed average F1-scores of 96.68%. The F1-score shows how well precision and recall are balanced; thus, both false negatives and positives are considered in this score. It is a crucial metric, especially if the dataset is imbalanced.

The computational time of our model was 6.845 seconds, which is less expensive than [28]. The AUC curve of our model based on the ten-fold CV is shown in

Figure 4. Figure 5 depicts the recall-precision curve. Figure 6 shows a threshold plot of the proposed model. Finally, we compared the suggested technique to state-of-the-art ML methods (RF and SVM) to further assess its performance and resilience. Accuracy, AUC, recall, Precision, Kappa, F1-score, and Mathew's correlation coefficient (MCC) were used to measure performance (see Table 4). We compared these models with our proposed model on the same study dataset.

**Table 3** Performance evaluation of our proposed model based on ten-fold CV

| Fold No. | Accuracy | AUC | Recall | Prec. | F1-score | Kappa | MCC |
|---|---|---|---|---|---|---|---|
| **0** | 0.9569 | 0.9932 | 1 | 0.9206 | 0.9586 | 0.9137 | 0.9171 |
| **1** | 0.9695 | 0.9994 | 0.9898 | 0.9426 | 0.9704 | 0.9391 | 0.9408 |
| **2** | 0.9695 | 0.9976 | 1 | 0.9426 | 0.9704 | 0.9391 | 0.9408 |
| **3** | 0.9632 | 0.9993 | 0.9797 | 0.9314 | 0.9645 | 0.9264 | 0.9289 |
| **4** | 0.9721 | 0.9990 | 0.9873 | 0.9470 | 0.9728 | 0.9442 | 0.9456 |
| **5** | 0.9720 | 0.9979 | 0.9873 | 0.9470 | 0.9728 | 0.9441 | 0.9456 |
| **6** | 0.9530 | 0.9963 | 0.9975 | 0.9140 | 0.9550 | 0.9060 | 0.9100 |
| **7** | 0.9581 | 0.9985 | 0.9975 | 0.9225 | 0.9597 | 0.9161 | 0.9194 |
| **8** | 0.9771 | 0.9928 | 0.9898 | 0.9562 | 0.9776 | 0.9543 | 0.9553 |
| **9** | 0.9644 | 0.9989 | 0.9949 | 0.9335 | 0.9656 | 0.9289 | 0.9312 |
| **Mean** | **0.9656** | **0.9973** | **0.9924** | **0.9357** | **0.9668** | **0.9312** | **0.9335** |
| **SD** | 0.0074 | 0.0023 | 0.0063 | 0.0129 | 0.0069 | 0.0148 | 0.0139 |

The AUC curve is a helpful and crucial estimate of general presentation and a generic measure of the classification model. The higher the AUC curve, the better the model's performance. The proposed approach's AUC curve was closer to the top-left corner of the figure, indicating that is suitable for early stroke detection. Overall, these findings support our proposed approach's success. The precision and recall curve is also frequently utilized to compare classifiers regarding recall and accuracy. The precision-recall curve gives a comprehensive view of categorization performance.

Our proposed technique attained the highest AUC score (99.73%), followed by the RF algorithm (77.53%), as shown in Table 4. The AUC score for the SVM linear method was the lowest (64.2%). Table 3 demonstrates that our method achieved the highest accuracy (96.56%), whereas the SVM algorithm had the lowest (59.02%). The proposed method attained a precision of approximately 93.57%. Furthermore, our method had the highest F1-score (96.68%), while the SVM algorithm had an F1-score of 60.74%, slightly outperforming the RF (49.06%).

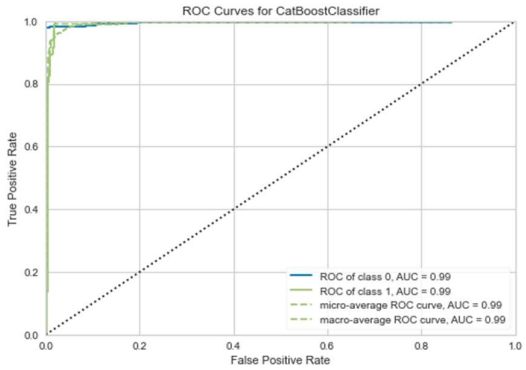


**Figure 4** Proposed method's AUC curve based on ten-fold CV.

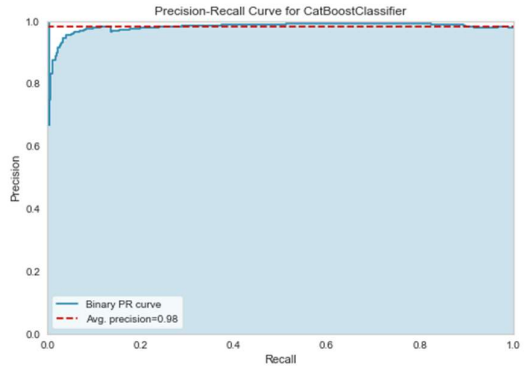


**Figure 5** Proposed method's precision-recall curve based on ten-fold CV.

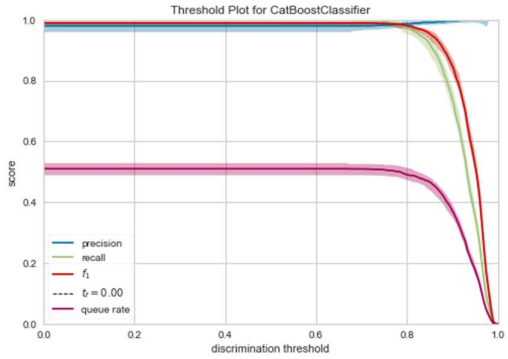


**Figure 6** Threshold plot of the proposed model.

Figure 7 shows the validation curve: the proposed model's training and CV scores. From Figure 7, the generalization gap of the proposed model is very moderate.

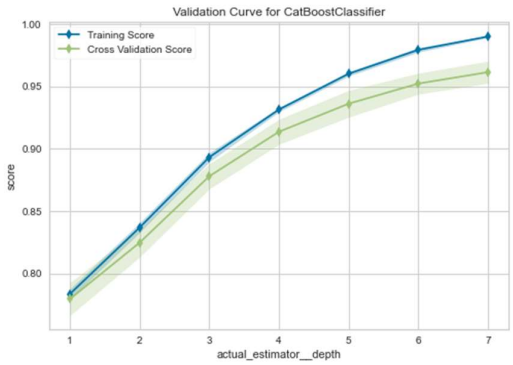


**Figure 7** Validation curve of the proposed approach.

**Table 4** Comparing the performance of the suggested technique to the SVM and RF machine learning algorithms.

| Model | Accuracy | AUC | recall | Prec. | F1 | Kappa | MCC |
|---|---|---|---|---|---|---|---|
| RF | 0.8374 | 0.7753 | 0.5833 | 0.5833 | 0.4906 | 0.527 | 0.5411 |
| SVM | 0.5902 | 0.642 | 0.64 | 0.5832 | 0.6074 | 0.6074 | 0.5827 |
| Proposed approach | **0.9656** | **0.9973** | **99.24** | **0.9357** | **0.9668** | **0.9312** | **0.9335** |

Our model's precision-recall curve was closest to the top-right corner (see Figure 5), indicating that the proposed approach performed better. The findings show that the proposed model outperformed the other classification techniques. One tree is built at a time using the selected method (Categorical Gradient Boosting Machine), with each new tree helping to fix mistakes caused by prior trained trees. While using a random input sample, the RF trains each tree separately. A tiny adjustment to the hyperparameters will impact practically all the forest's trees, which could influence the prediction outcome. Additionally, Random Forest strives to provide greater preference to hyperparameters to optimize the model, but Gradient Boosting Machine always pays more significance to functional space. Again, based on the well-known preprocessing technique known as target encoding, CatBoost is a novel approach of processing categorical characteristics that make this application more suitable. The findings also demonstrate the essence and worth of using an effective parameter optimization strategy to improve the prediction performance of the CatBoost model. Table 5 compares our proposed model and the literature using the accuracy metric. The proposed method performed moderately compared with the literature. Although

the accuracy performance of our model (0.94%) was lower compared with that of Lau, *et al.* [21], the computational complexity of our approach in training was 46.74% smaller. Thus, the proposed approach is less resource intensive.

**Table 5** Performance evaluation of our proposed methodology compared to alternative methods based on the accuracy metric.

| Approach | Accuracy |
|---|---|
| RF [20] | 71.6% |
| SVM [21] | 97.5% |
| DT, RF and MLP [25] | 74% - 75% |
| DT, MLP and Jrip [28] | 95% |
| Naive Bayes, Neural Networks, and DT [29] | 95% |
| ANN [31] | 89% |
| LightGBM [32] | 77.2% |
| Proposed approach | 96.56% |

## 5 Conclusions

In this work, we proved that integrated machine learning techniques can be used to forecast the likelihood of a patient getting a stroke. Early stroke identification is critical for achieving sustainable development goals (SDGs), specifically SDG 3. With the recent global increase in stroke prevalence and the rising difficulty of recognizing stroke early in developing countries like Ghana, a more efficient stroke detection method is critical to assist individuals and clinicians. This research uses an efficient categorical gradient boosting machine (CatBoost) to provide a smart strategy for early stroke detection. Several tests were carried out utilizing real-world datasets. In addition, the proposed technique was compared to previous study findings and state-of-the-art ML algorithms, i.e., RF and SVM. The empirical results showed that the proposed method beat previous studies, achieving the greatest accuracy (96.56%), Kappa (93.12%), AUC (99.73%), MCC (93.35%), precision (93.57%), and F1-score (96.68%). The findings also emphasize the relevance and usefulness of using an effective parameter optimization technique to improve the proposed approach's predictive performance.

## References

[1] Hilbert, A., Ramos, L.A., van Os, H.J.A., Olabarriaga, S.D., Tolhuisen, M.L., Wermer, M.J.H., Barros, R.S., van der Schaaf, I., Dippel, D., Roos, Y.B.W.E.M., van Zwam, W.H., Yoo, A.J., Emmer, B.J., Lycklama À Nijeholt, G.J., Zwinderman, A.H., Strijkers, G.J., Majoie, C.B.L.M. & Marquering, H.A., *Data-Efficient Deep Learning of Radiological Image Data for Outcome Prediction after Endovascular Treatment of Patients*

*with Acute Ischemic Stroke*, Comput. Biol. Med., **115**, 103516, 2019. DOI: 10.1016/j.compbiomed.2019.103516.

[2] Sirsat, M.S., Fermé, E. & Câmara, J., *Machine Learning for Brain Stroke: A Review*, J. Stroke Cerebrovasc. Dis., **29**(10), 105162, 2020. DOI: 10.1016/j.jstrokecerebrovasdis.2020.105162.

[3] Qiu, W., Kuang, H., Teleg, E., Ospel, J.M., Il Sohn, S., Almekhlafi, M., Goyal, M., Hill, M.D., Demchuk, A.M. & Menon, B.K., *Machine Learning for Detecting Early Infarction in Acute Stroke with Non–Contrast-enhanced CT*, Radiology., **294**, pp. 638-644, 2020. DOI: 10.1148/radiol.2020191193.

[4] Donkor, E.S., Owolabi, M.O., Bampoh, P., Aspelund, T. & Gudnason, V., *Community Awareness of Stroke in Accra, Ghana*, BMC Public Health, **14**, 196, 2014.

[5] Kashi, S., Polak, R.F., Lerner, B., Rokach, L. & Levy-Tzedek, S., *A Machine-Learning Model for Automatic Detection of Movement Compensations in Stroke Patients*, IEEE Trans. Emerg. Top. Comput., **9**(3), pp. 1234-1247, 2021. DOI: 10.1109/TETC.2020.2988945.

[6] Chang, C.Y., Cheng, M.J. & Ma, M.H.M., *Application of Machine Learning for Facial Stroke Detection*, in: Int. Conf. Digit. Signal Process. DSP, IEEE, pp. 1-5 , 2019. DOI: 10.1109/ICDSP.2018.8631568.

[7] Guberina, N., Dietrich, U., Radbruch, A., Goebel, J., Deuschl, C., Ringelstein, A., Köhrmann, M., Kleinschnitz, C., Forsting, M. & Mönninghoff, C., *r fvc of Early Infarction Signs with Machine Learning-Based Diagnosis by Means of the Alberta Stroke Program Early CT Score (ASPECTS) in the Clinical Routine*, Neuroradiology. 60, pp. 889–901, 2018. DOI: 10.1007/s00234-018-2066-5.

[8] Feigin, V.L., Forouzanfar, M.H., Krishnamurthi, R., Mensah, G.A., Connor, M., Bennett, D.A., Moran, A.E., Sacco, R.L., Anderson, L., Truelsen, T., O'Donnell, M., Venketasubramanian, N., Barker-Collo, S., Lawes, C.M.M., Wang, W., Shinohara, Y., Witt, E., Ezzati, M., Naghavi, M. & Murray, C., *Global and Regional Burden of Stroke during 1990-2010: Findings from the Global Burden of Disease Study 2010*, Lancet, **383**, pp. 245–255, 2014. DOI: 10.1016/S0140-6736(13)61953-4.

[9] Akinyemi, R.O. & Brainin, M., *The African Stroke Organization - A New Dawn for Stroke in Africa*, Nat. Rev. Neurol., **17**, pp. 127-128, 2021. DOI: 10.1038/s41582-021-00456-1.

[10] Ezejimofor, M.C., Uthman, O.A., Maduka, O., Ezeabasili, A.C., Onwuchekwa, A.C., Ezejimofor, B.C., Asuquo, E., Chen, Y.F., Stranges, S. & Kandala, N.-B., *Stroke Survivors in Nigeria: A Door-To-Door Prevalence Survey from the Niger Delta Region*, J. Neurol. Sci., **372**, pp. 262-269, 2017. DOI: 10.1016/j.jns.2016.11.059.

[11] Owolabi, M., Akarolo-Anthony, S., Akinyemi, R., Arnett, D., Gebregziabher, M., Jenkins, C., Tiwari, H., Arulogun, O., Akpalu, A.,

Sarfo, F., Obiako, R., Owolabi, L., Sagoe, K., Melikam, S., Adeoye, A., Lackland, D. & Ovbiagele, B., *The Burden of Stroke in Africa: A Glance at the Present and a Glimpse Into The Future: Review Article*, Cardiovasc. J. Afr., **26**, pp. S27–S38, 2015. DOI: 10.5830/CVJA-2015-038.

[12] Sampane-Donkor, E., *A Study of Stroke in Southern Ghana : Epidemiology*, *Quality of Life and Community Perceptions,* Doctoral Dissertation, Reykjavík: University of Iceland, School of Health Sciences, 2014.

[13] Akinyemi, R.O., Owolabi, M.O., Ihara, M., Damasceno, A., Ogunniyi, A., Dotchin, C., Paddick, S.-M., Ogeng'o, J., Walker, R. & Kalaria, R.N., *Stroke, Cerebrovascular Diseases and Vascular Cognitive Impairment in Africa*, Brain Res. Bull. **145**, pp. 97-108, 2019. DOI: 10.1016/j.brainresbull.2018.05.018.

[14] Sarfo, F.S., Akpa, O., Ovbiagele, B., Akpalu, A., Wahab, K., Komolafe, M., Obiako, R., Owolabi, L., Osaigbovo, G.O., Jenkins, C., Ogbole, G., Fakunle, A., Tiwari, H.K., Arulogun, O., Arnett, D.K., Asowata, O., Ogah, O., Akinyemi, R.O. & Owolabi, M.O., *Influence of Age on Links Between Major Modifiable Risk Factors and Stroke Occurrence in West Africa*, J. Neurol. Sci., **428**, 117573, 2021. DOI: 10.1016/j.jns.2021.117573.

[15] Sarfo, F.S. & Ovbiagele, B., *Prevalence and Predictors of Statin Utilization among Patient Populations at High Vascular Risk in Ghana*, J. Neurol. Sci., **414**, 116838, 2020. DOI: 10.1016/j.jns.2020.116838.

[16] Sailasya, G. & Kumari, G.L.A., *Analyzing the Performance of Stroke Prediction using ML Classification Algorithms*, Int. J. Adv. Comput. Sci. Appl., **12**, pp. 539-545, 2021. DOI: 10.14569/IJACSA.2021.0120662.

[17] Emon, M.U., Keya, M.S., Meghla, T.I., Rahman, M.M., Al Mamun, M.S. & Kaiser, M.S., *Performance Analysis of Machine Learning Approaches in Stroke Prediction*, in: 2020 4th Int. Conf. Electron. Commun. Aerosp. Technol., IEEE, pp. 1464-1469, 2020. DOI: 10.1109/ICECA49313. 2020.9297525.

[18] Akyeramfo-Sam, S., Addo Philip, A., Yeboah, D., Nartey, N.C. & Kofi Nti, I., *A Web-Based Skin Disease Diagnosis Using Convolutional Neural Networks*, Int. J. Inf. Technol. Comput. Sci., **11**, pp. 54-60, 2019. DOI: 10.5815/ijitcs.2019.11.06.

[19] Nti, I.K., Nyarko-Boateng, O. & Aning, J., *Performance of Machine Learning Algorithms with Different K Values in K-fold CrossValidation*, Int. J. Inf. Technol. Comput. Sci., **13**, pp. 61-71, 2021. DOI: 10.5815/ijitcs.2021.06.05.

[20] Menchón-Lara, R.M. & Sancho-Gómez, J.L., *Fully Automatic Segmentation of Ultrasound Common Carotid Artery Images Based on Machine Learning*, Neurocomputing, **151**, pp. 161-167, 2015. DOI: 10.1016/j.neucom.2014.09.066.

[21] Lau, H., Tong, K. & Zhu, H., *Support Vector Machine for Classification of Walking Conditions of Persons after Stroke with Dropped Foot*, Hum. Mov. Sci., **28**, pp. 504-514, 2009. DOI: 10.1016/j.humov.2008.12.003.

[22] Khosla, A., Cao, Y., Lin, C.C.Y., Chiu, H.K., Hu, J. & Lee, H., *An Integrated Machine Learning Approach to Stroke Prediction*, Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., pp. 183-191, 2010. DOI: 10.1145/1835804.1835830.

[23] Singh, M.S., Choudhary, P. & Thongam, K., *A Comparative Analysis for Various Stroke Prediction Techniques*, in Int. Conf. Comput. Vis. Image Process., Springer, Singapore, pp.98-106, 2019.

[24] Bandi, V., Bhattacharyya, D. & Midhunchakkravarthy, D., *Prediction of Brain Stroke Severity Using Machine Learning*, Rev. d'Intelligence Artif. **34**, pp. 753-761, 2020. DOI: 10.18280/ria.340609.

[25] Nwosu, C.S., Dev, S., Bhardwaj, P., Veeravalli, B. & John, D., *Predicting Stroke from Electronic Health Records,* Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. EMBS., pp. 5704-5707, 2019. DOI: 10.1109/EMBC.2019.8857234.

[26] Nti, I.K., Adekoya, A.F. & Weyori, B.A., *A Novel Multi-Source Information-Fusion Predictive Framework Based on Deep Neural Networks for Accuracy Enhancement in Stock Market Prediction*, J. Big Data, **8**, 17, 2021. DOI: 10.1186/s40537-020-00400-y.

[27] Nti, I.K., Adekoya, A.F. & Weyori, B.A., *A Systematic Review of Fundamental and Technical Analysis of Stock Market Predictions*, Artif. Intell. Rev., **53**, pp. 3007-3057, 2019. DOI: 10.1007/s10462-019-09754-z.

[28] Almadani, O. & Alshammari, R., *Prediction of Stroke using Data Mining Classification Techniques*, Int. J. Adv. Comput. Sci. Appl. **9**(1), pp. 453-460, 2018. DOI: 10.14569/IJACSA.2018.090163.

[29] Kansadub, T., Thammaboosadee, S., Kiattisin, S. & Jalayondeja, C., *Stroke Risk Prediction Model Based on Demographic Data*, in 8th Biomedical Engineering International Conference (BMEiCON), 2015.

[30] Jamthikar, A., Gupta, D., Saba, L. Khanna, N.N., Araki, T., Viskovic, K., Mavrogeni, S. Laird, J.R., Pareek, G., Miner, M., Sfikakis, P.P., Protogerou, A., Viswanathan, V., Sharma, A., Nicolaides, A., Kitas, G.D. & Suri, J.S., *Cardiovascular/Stroke Risk Predictive Calculators: A Comparison between Statistical and Machine Learning Models*, Cardiovasc. Diagn. Ther., **10**(4), pp. 919-938, 2020. DOI: 10.21037/cdt.2020.01.07.

[31] Shanthi, D., Sahoo, G. & Saravanan, N., *Designing an Artificial Neural Network Model for the Prediction of Thromboembolic Stroke*, Int. J. Biometrics Bioinforma, **3**(1), pp. 10-18, 2009.

[32] Kim, S.H., Jeon, E.T., Yu, S., Oh, K., Kim, C.K., Song, T.J., Kim, Y.J., Heo, S.H., Park, K.Y., Kim, J.M., Park, J.H., Choi, J.C., Park, M.S., Kim, J.T., Choi, K.H., Hwang, Y.H., Kim, B.J., Chung, J.W., Bang, O.Y., Kim,

G., Seo, W.K. & Jung, J.M., *Interpretable Machine Learning for Early Neurological Deterioration Prediction in Atrial Fibrillation-Related Stroke*, Sci. Rep., **11**, pp. 1-9, 2021. DOI: 10.1038/s41598-021-99920-7.

[33] Park, D., Jeong, E., Kim, H., Pyun, H.W., Kim, H., Choi, Y.-J., Kim, Y., Jin, S., Hong, D., Lee, D.W., Lee, S.Y. & Kim, M.C. *Machine Learning-Based Three-Month Outcome Prediction in Acute Ischemic Stroke: A Single Cerebrovascular-Specialty Hospital Study in South Korea*, Diagnostics, **11**(10), 1909, 2021. DOI: 10.3390/diagnostics11101909.

[34] Dorogush, A.V., Ershov, V. & Gulin, A., *Catboost: Gradient Boosting with Categorical Features Support*, ArXiv, abs/1810.11363, pp. 1-7, 2018.

[35] Huang, G., Wu, L., Ma, X., Zhang, W., Fan, J., Yu, X., Zeng, W. & Zhou, H., *Evaluation of Catboost Method for Prediction of Reference Evapotranspiration in Humid Regions*, J. Hydrol, **574**, pp. 1029-1041, 2019. DOI: 10.1016/j.jhydrol.2019.04.085.

[36] Hancock, J. & Khoshgoftaar, T.M., *Medicare Fraud Detection using CatBoost*, Proc. - 2020 IEEE 21st Int. Conf. Inf. Reuse Integr. Data Sci. IRI 2020, pp. 97-103, 2020. DOI: 10.1109/IRI49571.2020.00022.

[37] Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A.V. & Gulin, A., *Catboost: Unbiased Boosting with Categorical Features*, Adv. Neural Inf. Process. Syst. 2018-Decem, pp. 6638–6648, 2018.

[38] Kang, P., Lin, Z., Teng, S., Zhang, G., Guo, L. & Zhang, W., *Catboost-Based Framework with Additional User Information for Social Media Popularity Prediction*, MM 2019 - Proc. 27th ACM Int. Conf. Multimed., pp. 2677-2681, 2019. DOI: 10.1145/3343031.3356060.

[39] Breiman, L., *Random Forest*, Mach. Learn, **45**, pp. 5-32, 2001. DOI: 10.1023/A:1010933404324.

[40] Vapnik, V.N., *An Overview of Statistical Learning Theory, IEEE Trans. Neural Networks*, **10**, pp. 988-999, 1999. 10.1109/72.788640.

[41] Horak, J., Vrbka, J. & Suler, P., *Support Vector Machine Methods and Artificial Neural Networks Used for the Development of Bankruptcy Prediction Models and their Comparison*, J. Risk Financ. Manag., **13**, 60, 2020. DOI: 10.3390/jrfm13030060.

[42] Nti, I.K., Nyarko-Boateng, O., Adekoya, F.A. & Weyori, B.A., *An Empirical Assessment of Different Kernel Functions on the Performance of Support Vector Machines*, Bull. Electr. Eng. Informatics., **10**(6), pp.3403-3411, 2021. DOI: 10.11591/eei.v10i6.3046.

[43] Sarfo, F.S., Agyei, M., Ogyefo, I., Opare-Addo, P.A. & Ovbiagele, B., *Factors Linked to Chronic Kidney Disease Among Stroke Survivors in Ghana*, J. Stroke Cerebrovasc. Dis., **30**, pp. 1-7, 2021. DOI: 10.1016/j.jstrokecerebrovasdis.2021.105720.