# A Machine Learning Approach for Stroke Differential Diagnosis by Blood Biomarkers

Fayroz F. Sherif [1] and Khaled S. Ahmed [2,*]

[1] Computers and Systems Department, Electronics Research Institute, Egypt
[2] Biomedical Department, Faculty of Engineering, Benha University, Egypt
Email: fayroz_farouk@gmail.com (F.F.S.); khaled.sayed@bhit.bu.edu.eg (K.S.A.)
*Corresponding author

***Abstract*—Stroke happens when a clot blocks the blood supply to a region of the brain (ischemic stroke) or when an artery ruptures or spills blood (hemorrhagic stroke). Seeking medical care after a stroke may increase one's chances of survival and reduce long-term brain damage. Neuroimaging helps determine who and how to treat, although it is costly, not always accessible, and may have contraindications. These constraints lead to these reperfusion treatments being underutilized. Using a blood biomarker panel capable of consistently differentiating between ischemic stroke and intracerebral hemorrhage might be very beneficial and straightforward to deploy. Therefore, this study describes a system to speed and improve stroke diagnosis. Using four machine learning algorithms: Support Vector Machine (SVM), Adaptive Neuro-Fuzzy Inference System (ANFIS), K-Nearest Neighbor (KNN), and Decision Tree (DT), we aim to find promising blood biomarker candidates for differential stroke diagnosis. A two-stage binary classifier model was created to classify the stroke group vs. the normal group and then categorize the instances allocated to the stroke group into ischemic and hemorrhagic groups. Our findings reveal that SVM is better than ANN, ANFIS, and DT for distinguishing strokes in Egyptian patients, according to our data. The most important blood features are Absolute (ABS) Neutro, Creatine Phosphokinase (CPK), Neutro/Neutrophils, and White Blood Cell (WBC) Count/Leukocytes laboratory tests that may serve as crucial and significant indications for stroke diagnosis. The selected characteristics and a two-stage binary classifier discriminated with higher accuracy (Ischemic and hemorrhagic patients). This method for identifying and classifying brain strokes was accurate, easy to use, and cost-effective.**

***Keywords*—machine learning, blood biomarker, stroke, hemorrhagic, ischemic, identification and classification**

## I. Introduction

Globally, stroke is the leading cause of disability and death. Over half of stroke survivors have a permanent disability [1]. The stroke must be diagnosed quickly in order to provide acute intervention within 3–6 h, which is the top limit of the treatment window for the greatest long-term results which leads that hospital design and planning is of great importance. Emergency department relation with other departments of hospital especially radiology and/or laboratory is very critical to aid the early diagnosis even with normal conditions or in case of infection [2–4]. Ischemic Stroke (IS) and Hemorrhagic Stroke (HS) are the two primary subtypes of stroke. Ischemic Stroke (IS) is responsible for 85% of all cases. Ischemic stroke is caused by a blockage in a brain artery, whereas Hemorrhagic Stroke (HS) is the result of a brain hemorrhage. Both kinds can cause long-term harm that affects cognition and movement, as well as vision and communication. Patients with strokes are at high risk for long-term brain impairment, consequences (including disability), and even death [5]. When comparing brain strokes, medical treatments are problematic since there are no obvious borders between stroke types. In clinical practice, it is crucial to be able to distinguish between different kinds of strokes and to know when the stroke first occurred. Improving patient outcomes in acute stroke involves prompt and accurate identification of stroke and its subtypes. Many researches have been presented in COVID-19 based on blood biomarkers [6–8], and viral sequences [9, 10]. Other researchers tried to determine stroke non-invasively [11], or via exploring the protein functions [12].

Even though the current imaging-based methods for detecting a stroke are quick, they are not very good at diagnosing the problem in a clinical setting. The major goal of neuroimaging in a patient with suspected ischemic stroke is to rule out the existence of other forms of central nervous system lesions and to discriminate between ischemic and hemorrhagic stroke [13].

Computed tomography CT scans are deemed adequately sensitive for identifying mass lesions, such as a brain mass or abscess, as well as detecting acute bleeding. However, CT scans may not be sensitive enough to identify an ischemic stroke, particularly if it is small, acute, or in the posterior fossa (i.e., brainstem and cerebellar regions) (i.e., brainstem and cerebellum areas). Computed tomography CT may be normal in individuals with a minor stroke, resulting in a poor sensitivity of 30%. In comparison, Magnetic Resonance Imaging (MRI) is more sensitive (sensitivity of >80%), but it can't be used on restless patients (20–79% of all stroke patients) [14].

Manuscript received November 18, 2022; revised January 29, 2023; accepted June 6, 2023; published January 3, 2024.

Although MRI scans have a higher resolution than CT scans, they are less accessible and more costly. Additionally, MRI scans cannot be conducted on individuals who have certain implanted devices (e.g., pacemakers) or who suffer from claustrophobia. If a patient is within the time window for acute stroke intervention, guidelines say to order an MRI scan if it can be done as quickly as a CT scan. If it can't be done as quickly, CT is the preferred test because acute stroke treatments shouldn't be put off until more detailed imaging is available if the patient's history and physical exam are consistent with acute stroke. Also some studies have been conducted on Alzheimer which may have stroke effect [15, 16].

In addition to current imaging modalities, stroke management needs a new diagnostic approach with high speed, accurate detection, and cost-effective assays. An additional strategy for diagnosing acute stroke during the first hours is to find blood biomarkers. These markers represent the body's reaction to the damage produced by various forms of stroke. Biomarkers include proteins, ribonucleic acids, lipids, and metabolites. Specific blood biomarkers capable of distinguishing ischemic stroke from hemorrhagic stroke and its imitators, recognizing major vascular blockage, and predicting stroke start time might speed diagnosis and boost eligibility for reperfusion therapy. A biomarker that could discriminate between hemorrhagic and ischemic strokes and the danger of recurrent bleeding would, in principle, allow the broad beginning of thrombolysis in the ambulance and preserve critical time and brain damage.

Currently, blood biomarkers have low performance. Although biomarkers have the potential to improve stroke diagnosis and management, there is currently no marker that has demonstrated sufficient sensitivity, specificity, rapidity, precision, or cost-effectiveness to be used in routine stroke management, highlighting the need for additional research. To enhance biomarker performance, more standardization of clinical, laboratory and statistical methods across centers is required.

In case-control studies comparing stroke patients to controls, the mean values of several biomarkers were significantly different. However, the range of values has substantial overlap, and when these biomarkers are evaluated in prospective research, they demonstrate minimal use. Several factors might explain this poor performance, some of which are attributable to the assays themselves and others to stroke as a clinical entity [17]. Given the diversity of ischemic stroke, a single biomarker may not be enough to capture the underlying complexity. Biomarkers have been explored for a variety of therapeutic uses (risk for the development of the disease; diagnosis; characterization of clinical severity; identifying ischemic penumbra; estimating the risk of progression or worsening; and outcome). Although several biomarkers have been linked to brain ischemia, they provide no extra information for the patient beyond what can be gleaned through clinical examination and neuroimaging. More complicated models, including simultaneous measurements of numerous biomarkers, were explored to solve these constraints. In several investigations, researchers examined up to 50 distinct biomarkers at the same time. Only a few biomarkers achieved a sensitivity and specificity of >90%. Even adding some basic demographic information, like age, gender, or the presence or absence of atrial fibrillation, did not affect performance [18].

Computer-aided methods and statistical analysis have increased the accuracy of the process and model. Since many of these tools depend on human participation or the creation of characteristics, they are computationally costly and lack generalizability. In contrast, machine learning algorithms may learn from hidden data and provide a high degree of adaptability. Nonetheless, they also have the issue of addressing handmade characteristics and being data specific. Therefore, it is necessary to build a strategy with several parameters to learn and acquire the essential aspects, reducing the amount of human work required. This study gives an overview of the current strategies for detecting ischemic and hemorrhagic strokes using statistical and machine learning approaches in the different modalities. A recent study [19] reviewed 177 research publications published between 2010 and 2021 to highlight the present state and problems of Computer-Assisted Diagnosis (CAD), Machine Learning (ML), and Deep Learning (DL) based algorithms for CT and MRI as primary modalities for stroke detection and lesion area segmentation [20]. The last part of that review paper talks about the current needs, preferred mode, and possible research topics in the field.

A research paper [21] was done at six Catalan Stroke Centers. Patients suspected of having a stroke were included during the first 6 h of symptom onset, and blood samples were taken upon admission. Immunoassays were used to test a panel of 21 biomarkers chosen from prior findings and the literature. However, their results showed that the examined biomarkers are insufficient for a precise differential diagnosis of stroke. The logistic regression model for the comparison of stroke and stroke mimics did not include any biomarkers, but clinical factors were included with an accuracy of 80.8%. Also, the predictive accuracy was 80.6% when comparing ischemic versus hemorrhagic strokes.

To the best of our knowledge, there were no machine learning approaches that have been used in combination with common laboratory blood tests to differentiate stroke patients, and there was no mention of ordinary blood test datasets either in relevant papers. In this study, we aim to investigate the accuracy and performance of four soft computing techniques (Adaptive Neuro-Fuzzy Inference System (ANFIS), Support Vector Machine (SVM), K-Nearest Neighbor (KNN), and Decision Tree (DT)) for differentiating between the stroke group and the normal group, and then categorize the instances allocated to the stroke group into ischemic and hemorrhagic groups. The proposed study introduced an efficient and robust system model based on low-cost blood samples to assist pathologists in diagnosing ischemic and hemorrhagic stroke accurately. The study depends on machine learning techniques because of the number of collected cases.

## II. Materials and Methods

### A. Sample Collection

Patients enrolled in the study were all collected from the one hospital with different building having one Hospital Information System (HIS) considering the same types of medical equipment and all the patients had been physically checked and diagnosed via specialists. The full medical laboratory tests and brain imaging were recorded for each patient according to the hospital protocol (considering the same parameters in imaging modality together with same laboratory limits). This was cross-sectional research that included 410 people separated into three groups. The group of ischemic stroke patients is ($n$ = 145), the Hemorrhagic stroke group is ($n$ = 64), and the control group is ($n$ = 201). Around 50% of the sample was male and had an average age of 45 years. The details of these cases are indicated in Table I. The performed workflow is indicated in Fig. 1, where Table II summarizes the types of blood sample tests used for this study.

TABLE I. Dataset Details Used in This Study

| Stroke | Male | Female | Total |
|---|---|---|---|
| Ischemic stroke | 87 | 58 | 145 |
| Hemorrhagic stroke | 30 | 34 | 64 |
| Normal | 109 | 92 | 201 |
| Total | 226 | 184 | 410 |

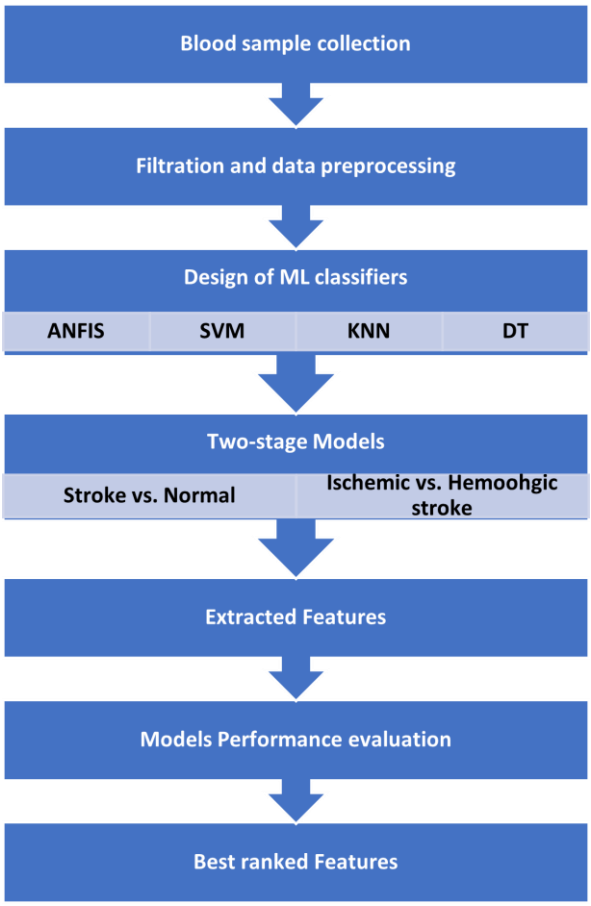


Fig. 1. A graphical representation of the procedure followed in this study.

TABLE II. The Types of Blood Tests Used in This Study

| No | Lab test abb. | Blood test Name |
|---|---|---|
| 1 | WBC Count | White Blood Cells |
| 2 | ANC | Absolute Neutrophil Count |
| 3 | Neutro | neutropenia |
| 4 | EOS | Eosinophils and Eosinophil Count Test |
| 5 | CR Protein | C-Reactive Protein (CRP) test |
| 6 | Creatinine | Creatinine levels in blood and/or urine |
| 7 | ABS BASO | Absolute Basophils count |
| 8 | MCHC | Mean Corpuscular Hemoglobin Concentration |
| 9 | Glucose R | Random glucose test |
| 10 | Urea | Blood Urea Nitrogen Test |
| 11 | K | Potassium serum Blood Test |
| 12 | CPK | Creatine Phosphokinase test |
| 13 | CKMB | Creatine Kinase-MB |
| 14 | Lymphocytes | Lymphocytes levels |
| 15 | Prothrombin | Prothrombin Time (PT) test |
| 16 | HCT ESR | The erythrocyte sedimentation rate |
| 17 | RBC Count | red blood cells |
| 18 | HGB | Hemoglobin Test |
| 19 | HCI | Hydrochloric acid (hydrogen chloride) |
| 20 | MCV | Mean Corpuscular Volume |
| 21 | MCH | mean corpuscular hemoglobin |
| 22 | RDW | Red Cell Distribution Width |
| 23 | Platelet count | Platelet count |
| 24 | MPV | mean platelet volume |
| 25 | ABS LYMPH | absolute lymphocyte |
| 26 | ABS MONO | Absolute (ABS) Monocytes |
| 27 | ABS EOS | Eosinophil count - absolute |
| 28 | MONO | Monocytes |
| 29 | CHOL | Cholesterol Levels |
| 30 | HDL | HDL cholesterol levels |
| 31 | LDL | LDL cholesterol levels |
| 32 | Albumin | Albumin Blood Test |
| 33 | Alkaline phos | Alkaline Phosphatase |
| 34 | ALT (SGPT) | Alanine Aminotransferase |
| 35 | Bilirubin | Bilirubin Test |
| 36 | TP | Total protein test |
| 37 | Globulin | Globulin Test |
| 38 | Ratio INR | Prothrombin (international normalized ratio) |
| 39 | CL | Chloride test |
| 40 | $CO_2$ | Amount of carbon dioxide in the blood serum |

### B. Data Preprocessing

The process of identifying, recognizing, and correcting faults in input data in order to reduce the influence of input data inaccuracies on subsequent studies is known as data preprocessing [22]. Some researchers have explored laboratory tests to be biomarkers for diagnosis, Management and treatment [23–25], for mortality Prediction Model [26], and as strok indicator [27]. The obtained laboratory datasets are often arranged as records containing dependent and independent data pairs and need preprocessing. In a dataset, data may be missing for one of two reasons: it may have never existed, or it may have been erased due to being deemed defective for some reason. The feature vectors that have many missing data points (>80%) were removed from the dataset. Using median imputation, missing values in continuous predictor variables were replaced. Moreover, the sparse laboratory data were also removed. We obtained fourteen filtered blood features out of forty that were included in our basic machine learning methods. Fig. 2 indicates the recommended workflow to diagnose the patients.

In supervised learning, the dataset consists of ($x$, $y$) pairs, where $x$ is a vector of input characteristics and y is the target output. Before usage in ML models, categorical variables were encoded. The processed data is further

separated into two categories: training and testing. The training datasets are used to train machine learning models. The test data is used to assess how successfully your model has been learned. In our study, 70% of our data is utilized for training, while 30% is preserved for testing. Here, four ML models will be applied and trained under two subsequent stages using ordinary laboratory blood tests. First, identifying the patients with stroke from other controls in which the models were trained on 147 samples as a positive cohort and 141 samples as a negative cohort. In the second stage, the models were trained on 102 samples as Ischemic Stroke (IS) and 45 samples as Hemorrhagic Stroke (HS). For preparation and analysis, open-source Python libraries were used.

### C. Model Development

As it was unknown which machine learning algorithms would perform best, we chose four common machine learning models with distinct algorithms to increase the probability of good discriminative performance: Adaptive Neuro-Fuzzy Inference System (ANFIS), Support Vector Machine (SVM), K-Nearest Neighbor (KNN), and Decision Tree (DT). Each approach was developed using the open-source Python libraries. In the present research, all the fourteen filtered blood features were included in our basic machine learning methods: ANFIS, SVM, KNN, and DT. In this section, the different methods of supervised learning are reviewed.

#### 1) Adaptive Neuro-Fuzzy Inference System (ANFIS)

ANFIS is a subclass of ANN that embodies the trade-off between ANN and fuzzy logic systems, combining smoothness provided by fuzzy control interpolation with adaptability provided by ANN back-propagation. ANFIS combines the advantages of ANN and fuzzy logic in a single implementation [28]. The ANFIS technique was employed in this study as implemented in Matlab's Fuzzy Logic toolbox, with a Sugeno-type Fuzzy Inference System (FIS) and Gaussian functions used to describe the fuzzy sets as membership functions. To model the training data, a hybrid learning algorithm was developed by integrating least-squares and back-propagation gradient descent techniques. ANFIS was ran 100 times for the sake of this research.

The architecture of the ANFIS model used in this research consists of an input layer, three hidden layers, and an output layer. The input layer contained fourteen input variables representing the 14 reduced blood features. The developed model's output variable was one binary neuron representing the "Normal" or "Stroke" case in the first stage and "ischemic "or "hemorrhagic" stroke in the second stage. ANFIS system consists of five phases:

**Fuzzification Phase** is the first phase. Herein, node *i* in this phase is a square node and is shown in Eq. (1).

$$O_i^1 = \mu_{A_i}(x), \quad for\ i = 1, 2 \tag{1}$$

where $x$ is the input to the $i^{\text{th}}$ node, $\mu_{A_i}(x)$ is the fuzzy Membership Function (MF). In this paper, the following Triangular MF is used.

$$\text{Triangular}(x; a, b, c) = \begin{cases} 0 & x \le a \\ \frac{x-a}{b-a} & a \le x \le b \\ \frac{c-x}{c-b} & b \le x \le c \\ 0 & c \le x \end{cases}$$

where $(x;\ a,\ b,\ c)$ is the parameter set that changes the shapes of the MFs. Parameters are the predictor in this layer.

**Product Phase:** The second phase in the ANFIS network is the rule layer where the membership functions are the input values, and each node multiplies the input and provides an output that reflects the rule's firing strength by multiplication. This layer's output is given in Eq. (2):

$$O_i^2 = w_i = \mu_{A_i}(x) * \mu_{B_i}(y), \quad for\ i = 1, 2 \tag{2}$$

**Normalized phase:** Here the $i^{\text{th}}$ node is calculated by the ratio of the $i^{\text{th}}$ rules firing strength to the sum of the rule's firing strengths.

$$O_i^3 = \overline{w}_i = \frac{w_i}{w_1 + w_2}, \quad \text{for i} = 1, 2 \tag{3}$$

where $\overline{w}_i$ is referred to the normalized firing strength.

**De-fuzzy phase**: The nodes in this layer are marked with a square, and they use the Eq. (4) to figure out the weighted output of each linear function. The output is calculated in this layer as the summation of all incoming signals.

$$O_i^4 = \overline{w}_t f_i = \overline{w}_t (p_i x + q_i y + r_i) \tag{4}$$

where $\overline{w}_i$ is the output of layer 3, and $\{p_i, qi,$ and $ri)$ are the Sugeno inference system's linear combination coefficients. these parameters are mentioned as the consequent parameters.

**Output phase:** The single node in this step collects the total output

$$O_i^5 = \sum_i \overline{w}_t f_i = \sum_i \frac{\overline{w}_i f_i}{w_i} = f_{out} \tag{5}$$

For estimating the premise and consequent parameters, ANFIS employs a hybrid learning technique. The technique of the hybrid learning algorithm estimates the subsequent parameters in a forward pass and the premise parameters in a backward pass. During the forward phase, the information propagates to Phase 4, when it is optimized using a least square regression approach. In the backward phase, error signals flow backward, and a Gradient Descent (GD) algorithm updates the premise parameters. Typically, this error measure is defined as the sum of the squared differences between measured and modeled values, and its value is reduced to a desirable level. The ANFIS output can be written as:

$$f_{out} = (\overline{w}_1 x)p_1 + (\overline{w}_1 y)q_1 + (\overline{w}_1)r_1 + (\overline{w}_2 x)p_2 + (\overline{w}_2 y)q_2 + (\overline{w}_2)r_2 \tag{6}$$

*2) Support Vector Machine (SVM)*

The Support Vector Machine (SVM) was developed by Vapnik [29]. It is a supervised learning approach from the area of machine learning theory and structural risk reduction that is suitable for both classification and regression. SVM utilizes training data to determine the maximum margin hyperplane that best divides data into groups or classes. The separation hyperplane is chosen such that it is the farthest away from the closest training data points of any class. The SVM's primary principle is to utilize a kernel function to project data in lower dimensional feature space (that may be nonlinearly separable) into points in a higher dimension space. The polynomial, Gaussian radial basis, and exponential Radial basis are examples of the kernels that are used in SVM to compute scores for each subject in a nonlinear issue. The data is then separated into classes (e.g., patient/normal) using an ideal hyperplane and the structural risk reduction principle [29].

The steps to perform SVM classifier:

1. Determine the class function from which the decision boundary will be selected. For linear SVM, the linear hyperplane equation is as follows:

$$w \cdot x + b = 0$$

2. Define the margin that consists of the minimum distance between a candidate's decision border and each class's point.

$$f(x) = \begin{cases} +1 & if \ \ w \cdot x + b \geq 0 \\ -1 & if \ \ w \cdot x + b < 0, \end{cases}$$

where $w$ and $b$ denote the vector (weight) of the regression coefficients and the intercept (bias) term, respectively.

3. Select the class's decision boundary (often the hyperplane) in Step (1).
4. Determine the effectiveness of the selected decision boundary on the training set.
5. Compute the predicted performance of categorization on the new data point.

*3) K-Nearest Neighbors (KNN)*

K-Nearest Neighbors (KNN) is a method for supervised classification. It classifies items by determining the distance between their unique feature values. It is non-parametric since it makes no assumptions about the distribution of the data, and it is also known as a lazy learning technique. It produces models without any training data points required. KNN utilizes existing data and classifies new data points using similarity metrics (e.g., the Euclidean distance, Manhattan distance, Cosine distance, or Hamming distance function) [30]. Classification is determined if a majority of k similar samples or the sample's nearest neighbors in the feature space belong to a certain category, then the sample must also belong to that category. KNNs are suitable for both classification and regression. In general, KNN may be executed with ease by following these steps:

1. Compute the distance metric between the new data and each sample from the training set.
2. Locate the k training sample that is closest to the new data.
3. Sort by the distance to the new data in decreasing order and choose the top *k* results.
4. Assign the new information to the predominant class.

Even though KNN doesn't make any assumptions, the data must be clean of outliers and samples with unclear categorization. In addition, the sizes of the classes should be approximately comparable to reduce bias when an unknown sample is assigned to a class. The simplest value of *k* is 1, however, it may be desirable to employ other numbers. If altering k-values results in changes to an object's categorization, the second option is not safe. In a more sophisticated version of this method, voting techniques other than the simple majority may be used, which may be advantageous if, for instance, the classes in the training set have vastly different variances.

*4) Decision Tree (DT)*

A decision tree employs a treelike graph to illustrate a flow-chart-like structure with a "root." Each tree node represents an attribute or group of attributes. The last node is a "leaf" that represents a class label. Both nominal and numerical input features are acceptable for decision trees. Decision trees are a non-parametric method; therefore, they don't make any assumptions about how space is distributed or how the classifier is built. There are several methods to create a decision tree from a dataset, depending on which characteristics to use for each node and what criteria to apply for splitting. The issue is to choose the best accessible qualities for each tree branch. "Best" optimizes information gain for that phase, hence the algorithm picks it. Information theory equations based on entropy are used to minimize entropy. In information theory, entropy measures randomness, not disorder. All-same items have a low entropy, whereas random items have high entropy. Minimizing entropy helps create decision trees.

$$E(S) \quad = \sum_{i=1}^{c} -p_i \log_2 p_i$$

where *E(S)* is the entropy of a dataset collection, *c* is the number of classes in the system, and $p_i$ is the percentage of instances that belong to class *i*.

*D. System Evaluation*

The criteria used to evaluate the machine learning model are crucial because they dictate how machine learning algorithm performance is assessed and compared. A confusion matrix is a table pattern that may be used to visualize the performance of a classification model. The performance of a classification model is evaluated using a N matrix, where N is the number of target classes. The matrix compares the actual target values to the classification model's forecast [31]. A two-by-two confusion matrix for binary or two-class classification has four outcomes: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). TP is the outcome when the model properly predicts the positive

class, while TN is the outcome when the model correctly predicts the negative class. FP and FN, on the other hand, are the inverse for TP and TN. The findings of the confusion matrix are used to create a variety of measurements. The most common criteria for measuring classification model performance are accuracy-focused, however, there are several alternative assessment measures available. In our study, the model's performance was assessed using measures such as Classification Accuracy (CA), Area under the Curve (AUC), F1-score, precision P, and recall R. Here, the performance of each of the four classifiers was evaluated and compared to assess the superior model in our two-stage binary classifier problem.

*1) Accuracy*

Accuracy is one of the most often used measures for evaluating classification performance; it is defined as the number of correct predictions provided by the model across all types of predictions. The benefit of accuracy is that it is simple to calculate with less complexity and simple for humans to grasp. When the target variable classes in the data are almost balanced, it suggests that the measurement approach is effective. However, apart from being unable to discriminate between the kind of errors it produces (FP vs FN), accuracy would not operate well if the data were skewed or unbalanced, and the use of additional assessment metrics should be explored [31].

$$CA = \frac{T_p + T_n}{T_p + T_n + F_p + F_n}$$

*2) Sensitivity*

Sensitivity, also known as the true positive rate, measures how successfully a classification system categorizes data points as positive. In some fields, sensitivity is usually combined with specificity to assess the prediction performance of a classification model or a diagnostic test. In binary classification, for example, sensitivity measures the proportion of positive instances obtained, while specificity measures the fraction of negative examples obtained. Sensitivity is defined as follows.

$$Sensitivity = \frac{T_p}{T_p + F_n}$$

*3) Specificity*

Specificity, also known as true negative rate, measures how successfully a classification system categorizes data points as negative. The following is a definition of specificity:

$$Specificity = \frac{T_n}{T_n + F_p}$$

*4) Precision*

Precision is the ratio of real positives to the total number of positives that a model predicts. In general, it verifies the accuracy of the forecasts. Precision is an excellent metric to use when the cost of a False Positive is significant when employing a classification model. Precision, like specificity, does not assess whether a binary classification's negative example is genuinely negative. According to [17] the definition of accuracy is provided below:

$$P = \frac{T_p}{T_p + F_p}$$

*5) Recall*

The recall is a statistic often used to pick the optimal model when the cost of false negatives is significant. In contrast to accuracy, specificity does not evaluate whether a negative example in a binary classification is genuinely negative.

$$R = \frac{T_p}{T_p + F_n}$$

*6) F1-score*

F1-score, often called F-measure, is a statistic that measures the harmonic mean of Recall and Precision values. It determines how many patterns in a certain class have been accurately detected. F1-score is preferable to accuracy when a balance between precision and recall is required and when class distribution is unbalanced. This suggests that a classifier with a high F1-score has both high accuracy and recall. In practice, there is typically a trade-off between precision and recall, such as increasing recall at the expense of precision by making the classifier more likely to make positive predictions and increasing precision at the expense of recall by making it less likely to make positive predictions The F1-score of a classifier is calculated as follows:

$$F1 - score = 2 \cdot \frac{Precision \times recall}{precesion + recall}$$

Specificity is the capacity of a biomarker to exclude the illness when it is not present. Sensitivity is the ability of a biomarker to identify the existence of a disease when the disease is present.

*7) The Area under the ROC Curve (AUC)*

Receiver Operating Characteristic (ROC) is the fundamental tool used in ROC analysis to address a variety of issues, including:

- Deciding a decision threshold that minimizes the mistake rate or misclassification cost for a certain class and the cost distribution.
- Locating a zone where one classifier performs better than another.
- Identifying places where classifiers perform less well than expected.
- Obtaining class posterior estimates that are calibrated.

The Area under the ROC Curve (AUC) is an essential and often used ROC curve-related ranking statistic. It was used to improve a learning model and to evaluate various learning algorithms. The AUC values reflected the overall ranking performance of a classifier. The AUC may be calculated for a binary issue as follows:

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

where $S_0$ is the sum of all the ranked positive (Class 0) examples, $n_0$ is the number of positive (Class 0)

examples, and $n_1$ is the number of negatives (Class 1) examples. Both in theory and practice, it has been shown that the AUC is a better way to measure classifier performance and find the best solution during classification training than the accuracy metric.

## III. Results and Discussion

Several image-based ML algorithms for stroke diagnosis have already been explored in the literature, and they have greatly improved the workflow of acute ischemic stroke patients. Most of these ML techniques aided in speedy stroke diagnosis and triaging, based on feature identification and segmentation. However, the reliable differential diagnosis for strokes is still difficult since it depends on so many patients' specific as well as clinical aspects. In addition, it is costly and not always accessible.

In our experiment, we utilized a collected dataset of blood ordinary tests to identify and classify patients with different stroke groups using different machine learning algorithms. ML models were applied and trained, under two subsequent stages. From the collected blood dataset, 141 blood samples without symptoms or abnormalities in the brain and 147 blood samples with abnormalities in the brain are extracted for this study. In the second stage, the models were trained on 102 samples as Ischemic Stroke (IS) and 45 samples as Hemorrhagic Stroke (HS). Python is utilized as an implementation tool to model the proposed brain stroke detection and classification technique for SVM, KNN, and DT. While Matlab's Fuzzy Logic toolbox is used to implement ANFIS model.

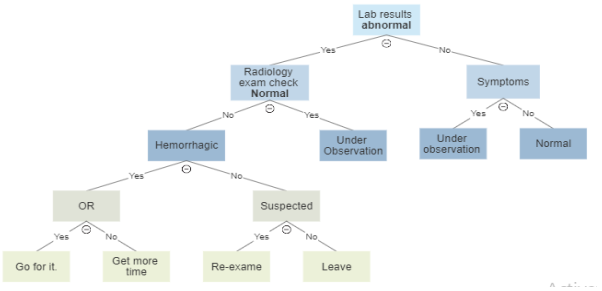


Fig. 2. A tress graphic indicates the patient tracking.

In the SVM algorithms, the Gaussian Kernel function's trade-off constant C is set at C = 0.1 and the parameter $\sigma$ of the Gaussian Kernel function was set to 0.02 in the SVM algorithms. In the KNN algorithm, the number of nearest neighbors was set to 5. The metric parameter was Euclidean distance, and the weights were uniform weights in which each neighborhood's scores are equally weighted. In the binary DT, the parameters are set and designed with a minimum number of splits equal to two (i.e., two child nodes) in the leaves and a maximum tree depth of 100 levels.

Assessing the performance of stroke vs. normal models was done by scoring the entire test sets with each model separately. Our findings demonstrate that SVM showed the best performance in discriminating between normal and stroke classes, achieving 100% in all performance metrics; as shown in Table III. The classification findings in terms of AUC, accuracy, F1-measure, precision, and recall of the first stage models were described in Table III. In the second stage classifiers, Table IV shows how well stroke groups (Ischemic vs. hemorrhagic) can be classified using all available ML methods and the fourteen based features.

TABLE III. Comparison of Performance Parameters (Stroke vs. Normal)

| Rank | Model | AUC | CA | F1 | Precision | Recall |
|---|---|---|---|---|---|---|
| 1 | SVM | 1.0 | 1.0 | 0.990 | 0.990 | 0.990 |
| 2 | ANFIS | 0.985 | 0.983 | 0.983 | 0.983 | 0.983 |
| 3 | KNN | 0.998 | 0.970 | 0.970 | 0.970 | 0.970 |
| 4 | Decision Tree | 0.975 | 0.985 | 0.985 | 0.985 | 0.985 |

TABLE IV. Comparison of Performance Parameters (Ischemic vs. Hemorrhagic Stroke)

| Rank | Model | AUC | CA | F1 | Precision | Recall |
|---|---|---|---|---|---|---|
| 1 | SVM | 0.991 | 0.991 | 0.991 | 0.991 | 0.991 |
| 2 | ANFIS | 0.97 | 0.967 | 0.970 | 0.970 | 0.97 |
| 3 | KNN | 0.98 | 0.982 | 0.982 | 0.983 | 0.982 |
| 4 | Decision Tree | 0.991 | 0.991 | 0.991 | 0.991 | 0.991 |

Table V ranks the features based on their association with a target variable, using appropriate internal scores (such as information gain, gain ratio, and chi-square) and ML model. The ranked five features; CPK, ABS Neutro, Neutro/Neutrophils, CKMB, and White Blood Cell (WBC) Count/Leukocytesare are efficient in discriminating between normal and stroke classes, achieving 100% accuracy. Additionally, Table VI ranked (descending sequence/order for parameter importance) the features according to how important they are in the classification models of Ischemic and Hemorrhagic stroke. Fig. 3 indicates the average, minimum and maximum limits of normal range and collected data for the most important parameters.

TABLE V. Ranking of Features in Classification of Stroke vs. Stroke

| Rank | Features | Info. gain | Gain ratio | Gini | ANOVA | $\chi^2$ | ReliefF |
|---|---|---|---|---|---|---|---|
| 1 | CPK | 0.524 | 0.262 | 0.259 | 146.61 | 88.96 | 0.391 |
| 2 | ABS Neutro | 0.434 | 0.217 | 0.224 | 62.40 | 78.59 | 0.219 |
| 3 | Neutro/Neutrophils | 0.231 | 0.115 | 0.109 | 1.337 | 9.92 | 0.115 |
| 4 | CKMB | 0.158 | 0.079 | 0.075 | 34.45 | 18.59 | 0.103 |
| 5 | WBC Count / Leukocytes | 0.154 | 0.077 | 0.061 | 25.51 | 20.49 | 0.087 |

TABLE VI. RANKING OF FEATURES IN THE CLASSIFICATION OF ISCHEMIC VS. HEMORRHAGIC STROKE

| Rank | Feature | Info. gain | Gain ratio | Gini | ANOVA | $\chi^2$ | ReliefF |
|---|---|---|---|---|---|---|---|
| 1 | ABS Neutro | 0.876 | 0.438 | 0.452 | 1000.2 | 70.74 | 0.54 |
| 2 | Neutro/Neutrophils | 0.518 | 0.259 | 0.288 | 50.57 | 47.917 | 0.233 |
| 3 | CPK | 0.524 | 0.262 | 0.265 | 106.52 | 50.89 | 0.196 |
| 4 | WBC Count / Leukocytes | 0.378 | 0.189 | 0.208 | 48.24 | 22.0 | 0.127 |

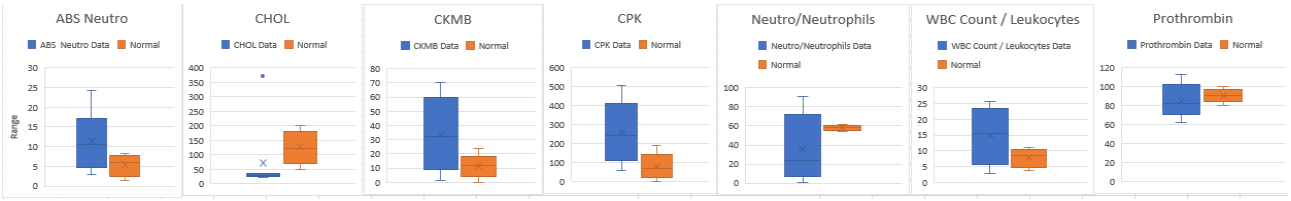


Fig. 3. Normal and abnormal ranges of most common parameters for stroke differentiation.

## IV. CONCLUSION

Improving patient outcomes in the acute phase of stroke needs prompt and accurate identification of stroke and its subtypes. The authors of this article used ML approaches to identify and classify stroke types using common laboratory blood tests. The ML approaches are validated in this study for classifying and predicting stroke subtypes, as well as obtaining blood feature reduction benefits. As a result, during real system operation, the suggested models employ just four features of the 14 stroke scale parameters, to offer quicker and more accurate service assistance. The best-ranked laboratory tests that varied considerably between the ischemic and hemorrhagic groups are (ABS Neutro, CPK, Neutro/Neutrophils, and WBC Count/Leukocytes). Our findings reveal that the SVM model is superior to KNN, ANFIS, and DT approaches for differentiating strokes in patients. However, future studies might expand on these results by examining if improved predictions can be made by utilizing a wider data set. Finally, the proposed study introduced an efficient and robust system model based on low-cost blood samples to assist pathologists in diagnosing ischemic and hemorrhagic stroke accurately.

## AVAILABILITY OF DATA AND MATERIAL

The data that support the findings of this study are available on request from the corresponding author.

## CONFLICT OF INTEREST

The authors declare no conflict of interest.

## AUTHOR CONTRIBUTIONS

F.F.S. and K.S.A. collected, analyzed, and interpreted the data. F.F.S. designed the work and implemented the software. K.S.A. supervised the implemented process. Both authors wrote the article and validated the results. Both authors had approved the final version.

## REFERENCES

[1] L. Feigin, B. Norrving, and G. A. Mensah, "Global burden of stroke," *Circ. Res.*, vol. 120, no. 3, pp. 439–448, Feb. 2017. doi: 10.1161/CIRCRESAHA.116.308413

[2] K. S. Ahmed, M. R. Ali, M. M. A. Lashin, and F. F. Sherif, "Designing a new fast solution to control isolation rooms in hospitals depending on artificial intelligence decision," *Biomedical Signal Processing and Control*, 2022.

[3] A. S. Elsaadany and A. Khaled "Adapting existing hospitals for infectious disease prevention and control," *Journal of Engineering Research*, vol. 6, pp. 6–12, 2022.

[4] K. S. Ahmed, "Medical planning: Operating theatre design and its impact on cost, area and workflow," in *Proc. Bioinformatics and Biomedical Engineering. IWBBIO 2017. Lecture Notes in Computer Science*, Springer, Cham, 2017, 10208. https://doi.org/10.1007/978-3-319-56148-6_28

[5] W. Xu *et al*., "Rapid computer-aided diagnosis of stroke by serum metabolic fingerprint based multi-modal recognition," *Adv. Sci*., vol. 7, no. 21, 2002021, Nov. 2020. https://doi.org/10.1002/advs.202002021

[6] D. R. Giacobbe, "Clinical interpretation of an interpretable prognostic model for patients with COVID-19," *Nature Machine Intelligence*, vol. 3, 16, 2021.

[7] M. Kermali, R. K. Khalsa, K. Pillai, Z. Ismail, and A. Harky, "The role of biomarkers in diagnosis of COVID-19—A systematic review," *Life Sci*., vol. 1, no. 254, 117788, 2020. doi: 10.1016/j.lfs.2020.117788

[8] H. Long, L. Nie, X. Xiang, H. Li, X. Zhang, X. Fu *et al*., "D-dimer and prothrombin time are the significant indicators of severe COVID-19 and poor prognosis," *BioMed Research International*, vol. 2020, 2020.

[9] F. F. Sherif and K. S. Ahmed, "Geographic classification and identification of SARS-CoV2 from related viral sequences," *International Journal of Biology and Biomedical Engineering*, no. 15, pp. 254–259, 2021.

[10] F. F. Sherif and K. S. Ahmed, "Unsupervised clustering of SARS-CoV-2 using deep convolutional autoencoder," *Journal of Engineering and Applied Science*, 2022

[11] A. El-Shahat and K. S. Ahmed, "Non-Invasive simulation technique for early stroke recognition using differentiation in magnetic flux density waves," *International Journal of Intelligent Systems and Applications in Engineering*, vol. 11, no. 2, pp. 854–863, 2023.

[12] K. S. Ahmed, N. H. Soloma, and Y. M. Kadah, "Exploring protein functions correlation based on overlapping proteins and cluster interactions," in *Proc. 2011 1st Middle East Conference on Biomedical Engineering*, Sharjah, United Arab Emirates, 2011, pp. 247–251.

[13] L. Herzog, E. Murina, O. Dürr, S. Wegener, and B. Sick, "Integrating uncertainty in deep neural networks for MRI based stroke analysis," *Medical Image Analysis*, vol. 65, 2020. doi: 10.1016/j.media.2020.101790

[14] Y. Mokli, J. Pfaff, D. P. Santos *et al*., "Computer-aided imaging analysis in acute ischemic stroke—Background and clinical applications," *Neurol. Res. Pract.*, vol. 1, no. 23, 2019. https://doi.org/10.1186/s42466-019-0028-y

[15] E. M. Arabi, K. S. Ahmed, and A. S. Mohra, "Advanced diagnostic technique for Alzheimer's disease using MRI top-

ranked volume and surface-based features," *J. Biomed Phys Eng*, 2022.
[16] E. M. Arabi, K. S. Ahmed, and A. S. Mohra, "High accuracy diagnosis for MRI imaging of Alzheimer's disease using XGBoost," *The Open Biotechnology Journal*, 2022
[17] R. Bhatia *et al*., "Role of blood biomarkers in differentiating ischemic stroke and intracerebral hemorrhage," *Neurol. India*, vol. 68, no. 4, pp. 824–829, 2020. doi: 10.4103/0028-3886.293467
[18] M. Dagonnier, G. A. Donnan, S. M. Davis, H. M. Dewey, and D. W. Howells, "Acute stroke biomarkers: Are we there yet?" *Frontiers in Neurology*, vol. 12, 2021. doi: 10.3389/fneur.2021.619721
[19] K. C. Ho, W. Speier, H. Zhang, F. Scalzo, S. El-Saden and C. W. Arnold, "A machine learning approach for classifying ischemic stroke onset time from imaging," *IEEE Transactions on Medical Imaging*, vol. 38, no. 7, pp. 1666–1676, July 2019. doi: 10.1109/TMI.2019.2901445.
[20] M. A. Inamdar *et al*., "A review on computer aided diagnosis of acute brain stroke," *Sensors*, vol. 21, no. 24. 2021. doi: 10.3390/s21248507
[21] A. Bustamante *et al*., "Blood biomarkers for the early diagnosis of stroke," *Stroke*, vol. 48, no. 9, pp. 2419–2425, Sep. 2017. doi: 10.1161/STROKEAHA.117.017076
[22] C. Sammut and G. I. Webb, *Encyclopedia of Machine Learning and Data Mining*, Springer, 2017. doi: 10.1007/978-1-4899-7687-1
[23] J. L. V. Reeve and P. J. Twomey, "Consider laboratory aspects in developing patient prediction models," *Nature Machine Intelligence*, vol. 3, no. 1, 18, 2021.
[24] J. Kamtchum-Tatuene and G. C. Jickling, "Blood biomarkers for stroke diagnosis and management," *Neuromol Med*., vol. 21, pp. 344–368, 2019. https://doi.org/10.1007/s12017-019-08530-0
[25] B. Pan, H. B. Alam, W. Chong *et al*., "CitH3: A reliable blood biomarker for diagnosis and treatment of endotoxic shock," *Sci. Rep*., vol. 7, 8972, 2017. https://doi.org/10.1038/s41598-017-09337-4
[26] N. N. Qomariyah, A. A. Purwita, S. D. A. Asri, and D. Kazakov, "A tree-based mortality prediction model of COVID-19 from routine blood samples," in *Proc. 2021 International Conference on ICT for Smart Society (ICISS)*, Bandung, Indonesia, 2021, pp. 1–7. doi: 10.1109/ICISS53185.2021.9533219.
[27] A. Gkantzios *et al*., "Evaluation of blood biomarkers and parameters for the prediction of stroke survivors' functional outcome upon discharge utilizing explainable machine learning," *Diagnostics*, vol. 13, 532, 2023. https://doi.org/10.3390/diagnostics13030532
[28] L. A. Zadeh, "Fuzzy sets as a basis for a theory of probability," *Fuzzy Sets Syst*., vol. 1, 1978.
[29] G. Castellazzi *et al*., "A machine learning approach for the differential diagnosis of Alzheimer and vascular dementia fed by MRI selected features," *Front. Neuroinform*., vol. 14, 2020. doi: 10.3389/fninf.2020.00025
[30] V. Vapnik and R. Izmailov, "Knowledge transfer in SVM and neural networks," *Ann. Math. Artif. Intell*., vol. 81, no. 1, pp. 3–19, 2017. doi: 10.1007/s10472-017-9538-x
[31] I. H. Sarker, "Machine learning: Algorithms, real-world applications, and research directions," *SN Comput. Sci*., vol. 2, no. 3, 160, 2021. doi: 10.1007/s42979-021-00592-x