Turkish Online Journal of Qualitative Inquiry (TOJQI)
Volume13, Issue 1, January 2022: 1349- 1356

Analysis and predicting the Wisconsin disease by using Machine Learning Algorithm

[1]Kumar R.G, [2]Bollineni Sree Chakrini, [3]Bandi Sai Bhanu Sekhar Reddy, [4]A.Suresh, [5]Shaik Basheer Ahmed, [6]Narra Sai Hari

[2,3,5,6] U G Student, [1,4]Associate Professor, [1,2,3,4,5,6] Department of Computer Science and Engineering, [1,2,3,4,5,6] Siddharth Institute of Engineering & Technology, Puttur-517583, India
Email: [1]rgkumarsietk@gmail.com, [2]bsreechakrini@gmail.com, [3]bandisaibhanu@gmail.com, [4]csesuresh6@gmail.com, [5]sbasheerahmed186@gmail.com, [6]saihari435@gmail.com.

Abstract

Computer vision should be widely used in medical applications, such as determining the sort of cancerous cells. Carcinoma is among the most frequent malignancies, and it kills a lot of people every year. It is the most frequent malignancy in females and the leading cause of death in women around the world. Cancerous cells were classified as either malignant (M) or benign (B) Some of the approaches used to classify and forecast carcinoma include the Decision Tree (CART), Naive Bayes (NB), k Nearest Neighbors (KNN)Support Vector Machine (SVM). A work, an (SVM) on the Wisconsin Carcinoma database was used. The dataset was additionally trained using the KNN, Naive Bayes, and CART methods, with predictive performance for each approach compared.

Keywords— KNN, Naive Bayes, CART, SVM, breast cancer

I. INTRODUCTION

Carcinoma was the most common malignant in women or has been the leading reason for mortality in women. Early diagnosis of malignant cells could help to reduce these deaths. To detect malignant cells, a variety of procedures should be utilized, including an MRI, mammography [1], sonography, and biopsies. In this research, the features were estimated to a capture picture of a fine needle aspiration (FNA) biopsy of a malignant tumor. Humans define the features of the image's cell nuclei. Breast cancer is diagnosed by categorizing the tumor. Tumors were divided into two categories: harmless and cancerous. Cancerous ones are more harmful than malignancies that are harmless. Nevertheless, not all clinicians are equipped to distinguish between benign and tumors, [2] as well as different malignancy units, might to two days. The correctly and effectively distinguish the type of malignancy units, computer vision approaches are used. Computer vision was a subset of deep learning (DL) which enables calculated to improve and advance with being pattern recognized. The purpose of machine learning was to create application software that could collect information and learn on its own. Some of the approaches used include Decision Tree (CART), Naive Bayes (NB), k Nearest Neighbors (k-NN), and Support Vector Machine (SVM) [3].

the training dataset was used to generate KNN suggestions. Forecasts were created by scanning the complete training database for the [4] K nearest neighbor's examples and integrating the outcome variable in each of those K instances for a new observation (x). The method could be (great common) class value in classification or the mean result parameter in extrapolation. A distance measure was utilized to estimate which of the K examples in the training data would be most similar to the new instance. Euclidean distance has been the most commonly used distance measure for real-valued input parameters. To calculate Euclidean distance, take the square root of the total of the diagonal deviations among a particular node (x) and current node (xi) overall outcome characteristics j.

$$Euclidean\ Distance\ (i, ix) = sqrt\big(sum((iy - ixy)^2)\big)$$

In a high-dimensional feature space, the attribute values were class-labeled matrices. The method's learning phase comprises solely of gathering the characteristic space and corresponding labels from supervised learning [5-8]. In the categorization, k should a utilized defined parameter, but also an unprocessed matrix (an inquiry or analysis domain) was categorized to select the term that occurs great frequently to the k learning instances which are closest to that query instance.

The Bayes' Principle would be utilized to develop the NB categories, which would be a set of categorization techniques. It's a collection of techniques that all operate on the same premise: every set of classified qualities was autonomous of the others. Bayes' Principle was P(h|d) = (P(d|h) * P(h)) / P(h) (d)

NB should be a categorization approach for binary (two-class) and multi-class categorization issues [9-12]. The approach was easy to comprehend when expressed used to binary or categorical input parameters. Suppose they should be a database for various categories of data. They had p1(h,d) of the likelihood to a bit of data belonging to Category 1 and p2(h,d) for the likelihood of a piece of information about Categorizes 2. (h,d). following criteria are used to categorize a new observation using parameters (h,d): the category was 1 if p1(h,d) was higher than p2(h,d). If p2(h,d) > p1, the category was (h,d).

(CART) would be a prediction representation that illustrates how the numbers of an outcome measure could be predicted using the values of other variables. A logistic regression would be the result of a CART, with every fork representing a split in a regression model and each final cluster representing a prediction for an outcome measure. The CART model was demonstrated using a binary tree. A different-input parameter (x) and a divided location of that parameter are provided by every parent node. (Let's pretend the parameter is an integer) Decision trees of the tree have an outcome parameter (y) that would be utilized to make a prediction.

SVM would be a deep learning method that could be utilized to solve categorization or regression issues. Regrettably, it is usually used to tackle classification problems. In this method, which depicts each information piece as a pixel in n-dimensional reality, the value of each feature has been the amount of a specific position (where n is several characteristics you have). Researchers next divide the data into two groups by determining the subspace that adequately divides them. Support Vectors are made up of single observation parameters [13-15]. (SVM) would be a threshold that best distinguishes between the 2 groups (hyper-plane/line). SVMs locate the subspace that separates the predictor variables into two classes after transferring the outcome sequence to a higher-dimensional feature space.

The minimal length between the chosen subspace and the incidences that are nearest to the border was achieved. A generated predictor has a high degree of generalization and can thus be used to reliably classify fresh samples. It's worth mentioning that SVMs could also produce probabilistic outcomes. The diagram below shows how an SVM could be used to determine whether a malignancy was benign characteristics such as size and the patient's age. The newly found subspace could be viewed as a boundary between two different groups. In general, the addition of a decision surface enables the discovery of every technique ambiguity.

II. LITERATURE SURVEY

[1] Was used to create the Wisconsin Carcinoma database. Contrasted to probabilistic classifications and Convolution Neural Networks, researchers suggested Support Vector Machines (SVMs) different classifiers to the prediction and diagnosis of malignancy disease in [2] The document contains operational specific as well as accompanying conclusions for each of the studied groups. An SVM model for screening mammography and prediction is developed in the Wisconsin Prognostic Breast Cancer (WPBC) or Wisconsin Diagnostic Breast Cancer (WDBC) databases. An optimized SVM technique performs well, with good accuracy (up to 96.91%), selectivity (up to 97.67%), and affectability (up to 96.91%). (up to 97.84 percent). According to [3] artificial neural networks are the extensively utilized forecasting methodology in medical prediction, despite their complex nature. The paper compares and contrasts the advantages and disadvantages of various machine learning approaches, including, the Naive Bayes Decision tree, neural networks, and SVM. Depending on the dataset and variable choice, each technique works differently in [4]. The KNN approach has produced the best outcomes in terms of the overall approach. NB and linear regression were accomplished to the diagnosis of carcinoma. However, SVM should be an important technique for predicting breast cancer reduplication and non-reduplication.

III. MATERIALS AND METHODS

The resources they used included Python coding programs and carcinoma information from the UCI repository. Our approach employs deep learning models to SVM, KNN, and Naive Bayes decision trees.

A. Dataset

The Wisconsin Diagnostic Melanoma dataset was collected from the University of California, Irvine's deep learning repository (http://archive.ics.uci.edu/ml). In a database, there are 357 benign melanoma cases and 212 aggressive melanoma patients. ID quantity would be the first units in a dataset, followed by the prognosis outcome (benign and malignant) and the average, sample variance, and mean of the worst readings, which are generally the third and fourth columns. There were no values that were missing. The parameters are derived from a computerized snapshot of the malignancy's fine needle aspiration endoscopy. The training information is a collection of data that is utilized to learn how to use the program. This information has been used to train the machine for performing various tasks.

B. Methodology

The database was split into two parts: learning and validation. The network is trained with 80% of the data, while the remaining 20% should be used for validation. Humans analyze the data and create a method to forecast whether a group of symptoms will progress to carcinoma. Computer learning algorithms were learning on training data before being tested on unlearning information. Overfitting is an issue that exists when a system was overly complex, such as when there are too many variables. Underfitting happens when a framework was overly basic and fails to capture the underlying trend of the data. Poor prediction accuracy was caused by both overfitting and underfitting. Cross-validation, regularisation, and dropout are three approaches for overcoming overfitting. K-fold cross-verification was among the most popular methods, in which the entire document was arbitrarily divided into the discriminate analysis of equivalent length. The model is tested with one sample group from the k, while the remaining

k-1 subsets were being utilized to build the machine. The k results are then combined to produce a single estimate. The fact that each testing subset was only utilized once is among the advantages of k-fold cross-validation shown in Figure 1. The (SVM) would be a binary classification that looks at the subspace for the largest possible proportion of pixels from the different categories on the different edge while optimizing the separation between every category and the subspace. SVMs should be a modern technique for cancer prediction and prognosis machine learning technologies. SVMs discover the subspace that splits the data points into 2 groups after transferring the support vector into a greater-dimensional feature set. The minimal distance between the chosen subspace and the incidences that were nearest to the border was maximized. The generated predictor has a high degree of generalization and could therefore be used to reliably classify fresh examples.

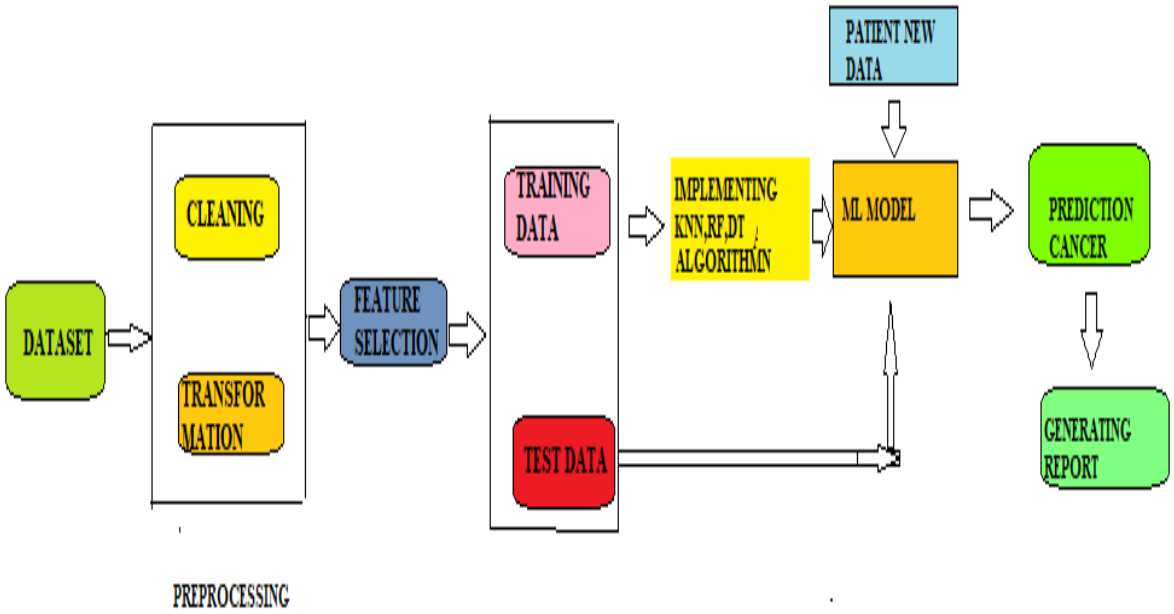

Figure 1: Proposed Architecture

C. Model selection

Controlled learning is an approach to teaching a machine of learning from information with clearly labeled input and output. The model is constructed from the learning data and applies what it has learned to determine the outcome of new information.

They're separated into three categories: extrapolation and categorization. A regression problem happens when the outcome is a realistic or constant variable, such as "salary" and "weight." It was a classification difficulty when the outcome is a category, such as "spam" and "not spam." The outcome variable or predictor variable, Y, in our database has just two sets of values: M (Malign) or B (Balance) (Benign). As a consequence, it is classified using a classification algorithm classification model. We've chosen three primary types of classification approaches in Computer Vision. To overcome problems, they might employ a simple little linear equation.

IV. RESULTS AND DISCUSSIONS

A database is created using the information gathered from the patients. The dataset is divided into learning and validation information, as well as there are no missing values in the database.

The four machine learning techniques were put to the test, and their predictive performance was evaluated. For forecasting, the method with the highest precision was utilized shown in Figure 2.

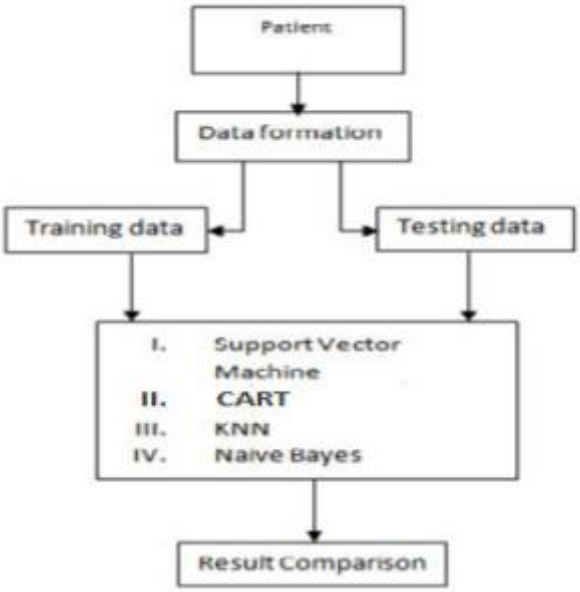

Figure 2: Flow diagram of various algorithms

A. Data Exploration

According to the ranges in the average, confidence deviation, and worst normal of the 10 features obtained from the concavity, compactness, fine needle aspiration slides geometrical parameters, softness, and parallelism all had surprisingly low scores to the evaluation. The quantity values for diameter, perimeter, and texture are all reasonably big, with areas that show the most volatility and the assessment value for all three variables. From the dispersal picture, humans could see that the malignant classification group has a higher mean for all of the characteristics.

B. Correlation

Humans could observe that several of the ten qualities' mean measurements were substantially connected. The red border around the diagonal denotes a correlation between characteristics. The yellow and green patches suggest a potential level of association, whereas the blue boxes indicate a negative association.

C. Count of Benign(B) and Malignant(M):

The total of MALIGNANT patients would be more than the total of BENIGN patients, as evidenced by the bar graph in Figures 3 -6 from our Jupiter notebook. Benign:357 Malignant:212

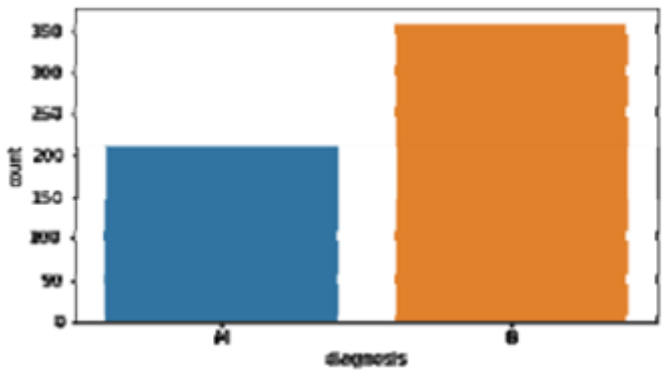

Figure 3: Patients count on types of cancer

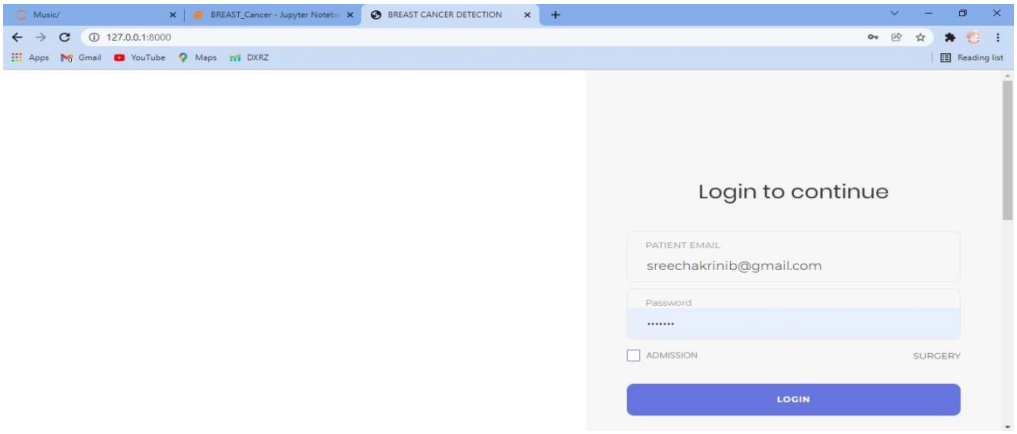

Figure 4: Login webpage

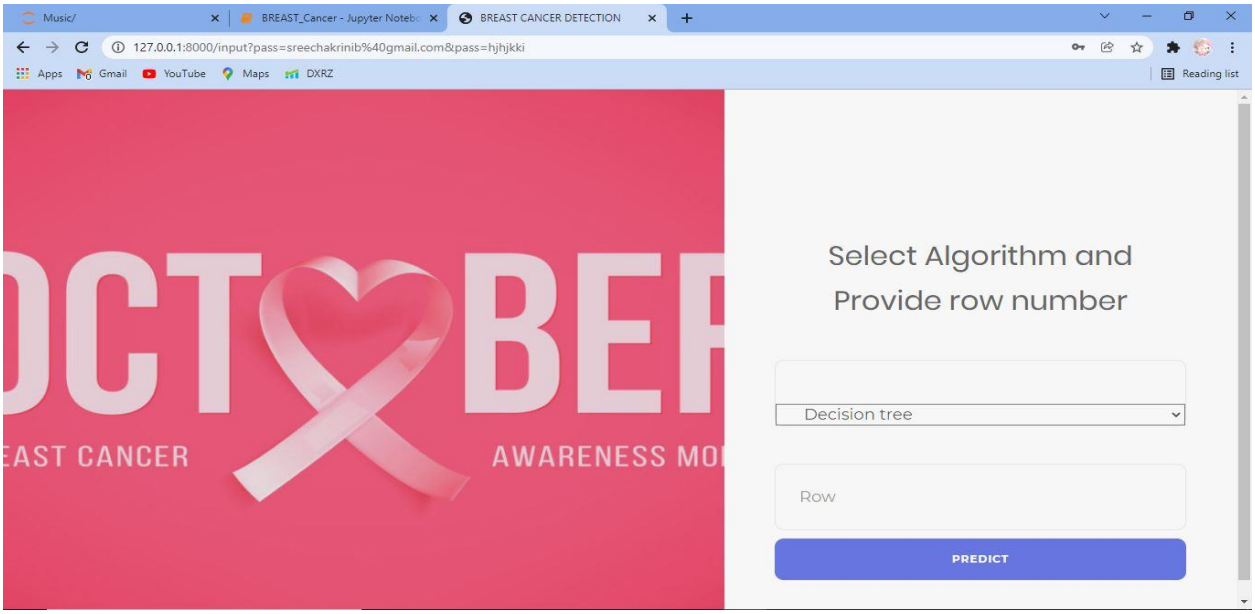

Figure 5: Choosing the algorithms

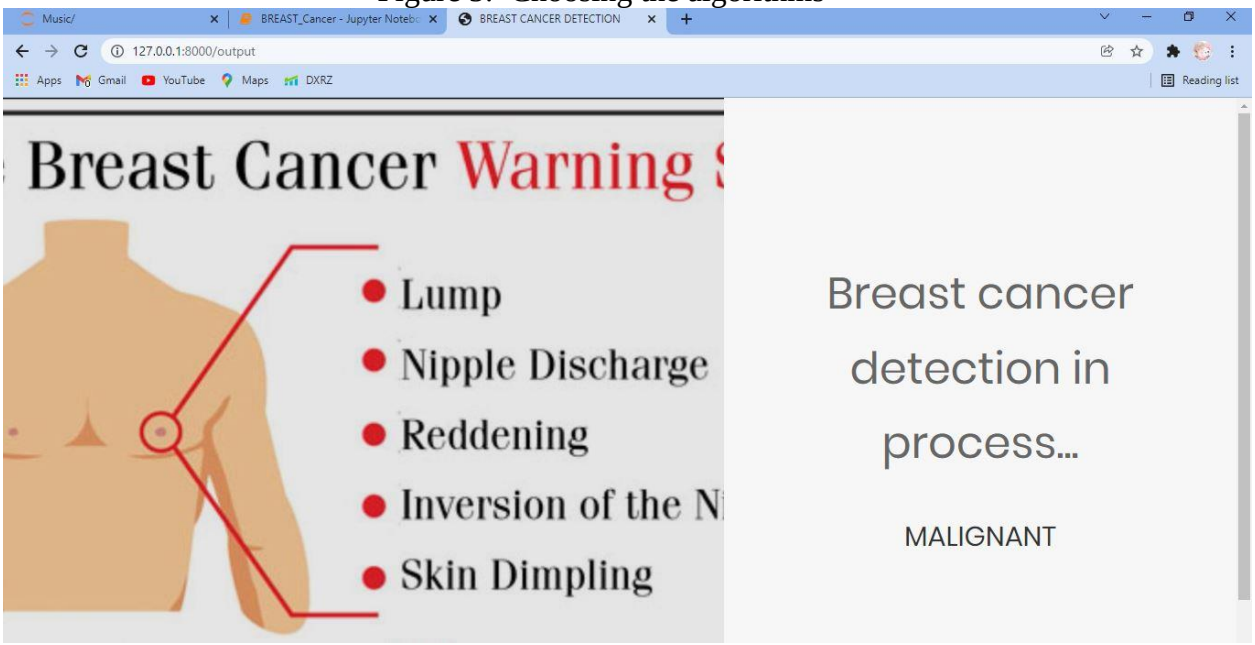

Figure 6: Performance analysis

D. Performance Comparison

According to the results, CART, KNN, Stochastic NB appear to outperform in the first round (all above 92 percent average accuracy). The SVM performs fairly poorly in this case. The effectiveness of the original dataset would increase if they standardize it.

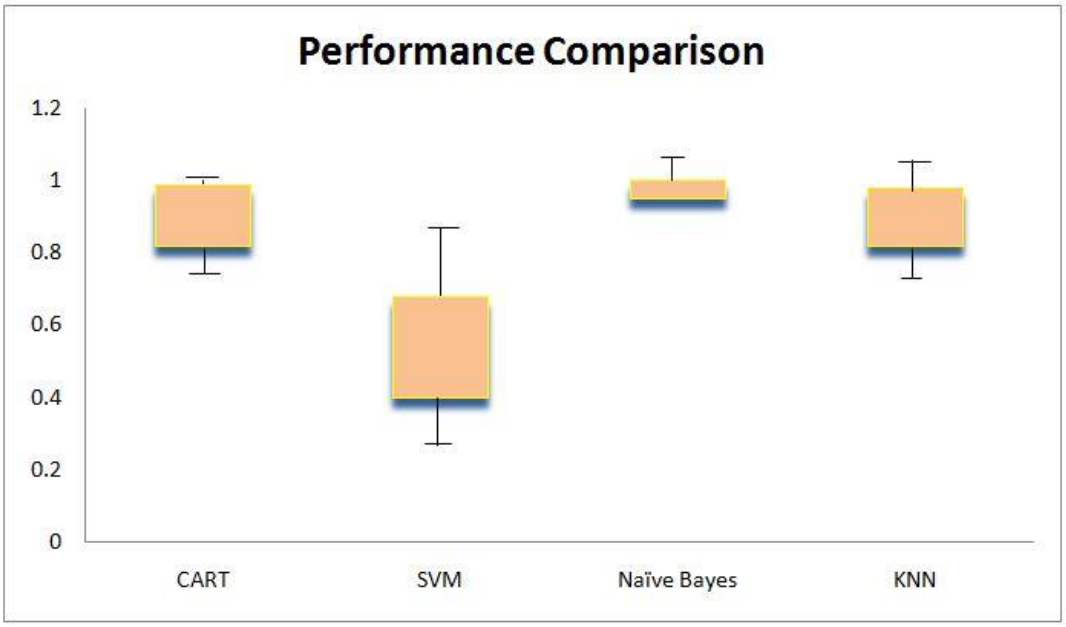

Figure 7: Comparison analysis

V. CONCLUSION AND FUTURE SCOPE

According to the dataset and the parameters selected, this method works individually. In terms of the total technique, the KNN technique has exhibited good results. Linear regression and Naive Bayes had also achieved well in the detection of carcinoma. SVM would be a powerful tool for forecasting validated and based on the findings, they believe that SVM with a Gaussian kernel would be the important method for predicting the recurrence of carcinoma. The SVM utilized in this study was still relevant when the number of group variables was binary, that when there are no more than two classes. Multiclass SVM was developed by scientists to address this issue. More research has been done in this area, including the construction of SVM classes like LIBSVM. The efficiency of methods could be improved by fine-tuning the parameters utilized in them. Moreover, for simplicity of use, this could be performed on a cloud service.

References

[1] Ali, S. F., & Padhi, R. (2011). Optimal blood glucose regulation of diabetic patients using single network adaptive critics. *Optimal Control Applications and Methods*, *32*(2), 196-214.

[2] Bamgbose, S. O., Li, X., & Qian, L. (2017, October). Closed loop control of blood glucose level with neural network predictor for diabetic patients. In *2017 IEEE 19th International Conference on e-Health Networking, Applications and Services (Healthcom)* (pp. 1-6). IEEE.

[3] Charles, R. K. J., Mary, A. B., Jenova, R., & Majid, M. A. (2019). VLSI design of intelligent, Self-monitored and managed, Strip-free, Non-invasive device for Diabetes mellitus patients to improve Glycemic control using IoT. *Procedia Computer Science*, *163*, 117-124.

[4] Garikapati, P., Balamurugan, K., Latchoumi, T. P., & Malkapuram, R. (2021). A Cluster-Profile Comparative Study on Machining AlSi7/63% of SiC Hybrid Composite Using Agglomerative Hierarchical Clustering and K-Means. Silicon, 13(4), 961-972.
[5] Aroulanandam, V. V., Latchoumi, T. P., Balamurugan, K., & Yookesh, T. L. (2020). Improving the Energy Efficiency in Mobile Ad-Hoc Network Using Learning-Based Routing. Rev. d'Intelligence Artif., 34(3), 337-343.
[6] Ranjeeth, S., Latchoumi, T. P., & Paul, P. V. (2020). Role of gender on academic performance based on different parameters: Data from secondary school education. Data in brief, 29, 105257.
[7] Venkata Pavan, M., Karnan, B., & Latchoumi, T. P. (2021). PLA-Cu reinforced composite filament: Preparation and flexural property printed at different machining conditions. Advanced Composite Materials, https://doi. org/10.1080/09243046.2021, 1918608.
[8] Ezhilarasi, T. P., Dilip, G., Latchoumi, T. P., & Balamurugan, K. (2020). UIP—a smart web application to manage network environments. In Proceedings of the Third International Conference on Computational Intelligence and Informatics (pp. 97-108). Springer, Singapore.
[9] Latchoumi, T. P., & Parthiban, L. (2017). Abnormality detection using weighed particle swarm optimization and smooth support vector machine. Biomedical Research, 28(11), 4749-4751.
[10] Aroulanandam, V. V., Latchoumi, T. P., Bhavya, B., & Sultana, S. S. (2019). Object detection in convolution neural networks using iterative refinements. architecture, 15, 17.
[11] Jernelv, I. L., Milenko, K., Fuglerud, S. S., Hjelme, D. R., Ellingsen, R., & Aksnes, A. (2019). A review of optical methods for continuous glucose monitoring. *Applied Spectroscopy Reviews*, *54*(7), 543-572.
[12] Zhang, J., Taniguchi, T., Takita, T., & Ali, A. B. (2003). A study on the epidermal structure of Periophthalmodon and Periophthalmus mudskippers with reference to their terrestrial adaptation. Ichthyological Research, 50(4), 310-317.
[13] Latchoumi, T. P., Balamurugan, K., Dinesh, K., & Ezhilarasi, T. P. (2019). Particle swarm optimization approach for waterjet cavitation peening. Measurement, 141, 184-189.
[14] Nahavandi, P. (2020). *Developing novel non-invasive MRI techniques to assess cerebrospinal fluid-interstitial fluid (CSF-ISF) exchange* (Doctoral dissertation, UCL (University College London)).
[15] Paul, M. C., Kir'yanov, A. V., Barmenkov, Y., Duarte, J., Leitão, J. P., Ferreira, M. F., ... & Dutta, D. A new class of specialty optical fibers based on a novel material composition of the doping host for the study of optical amplification.