2023 International Conference on Distributed Computing and Electrical Circuits and Electronics (ICDCECE)

2023 International Conference on Distributed Computing and Electrical Circuits and Electronics (ICDCECE) | 979-8-3503-4745-6/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICDCECE57866.2023.10151332

Information Fusion for Seveirty Detection using Machine Learning in Data Mining Healthcare

1st Vaishali Gupta
Department of Computer Science & Engineering,
Chandigarh Engineering College,
Jhanjeri, India
vaishali.j1631@cgc.ac.in

2nd Pawan Bhambu
Department of Computer Science Enginnering,
Vivekananda Global University,
Jaipur, India
pawan.bhambu@vgu.ac.in

***Abstract*—Improved patient care and less mental strain on healthcare providers are two benefits of using algorithms for machine learning to healthcare data. These algorithms may be used to spot irregularities in vital signs, which might speed up medical assistance or provide light on a disease's progression. While there is a wealth of literature comparing the unsupervised and supervised performances of anomaly detection algorithms on popular public datasets, this same level of conceptual comparison is lacking when it comes to physiological data. Knowing one's heart rate may provide valuable insight on one's health and level of physical activity, making it an underutilised data source. Specifically, we used and compared five machine learning methods, two of which were unsupervised and the other three supervised, to identify outliers in heart rate data. The algorithms were tested using physiological data from human subjects' hearts. Results demonstrated that both outlier factor and regression trees algorithms were effective in detecting heart rate anomalies, with both models successfully generalising from their simulation heart rate data training to real-world heart rate data. In addition, the findings lend credence to the idea that, in the absence of real labelled data, simulated data can be used to configure methodologies to a certain degree of performance, indicating that this kind of training could be particularly useful in the initial rollout of a system with no preexisting data.**

Keywords— Information Fusion, Severity Detection, Machine Learning, and Data Mining Healthcare

I. Introduction

The safe and continuing functioning of electrical and mechanical automation systems depends on accurate diagnosis of thermal defects in critical elements before they become severe. The overall management of thermal defects is not optimal due to the inability of current algorithms to create consistent links among sensors. In order to regulate the built environment in a way that minimises energy consumption while maximising efficiency, building's fault detection and diagnosis (FDD) technology is required. When dealing with complicated systems and unpredictable inputs, recent data-driven approaches have shown to be advantageous. Existing studies on data-driven FDD, however, treat the issue as nothing more than a classification challenge to label fault kinds. For a long time, prior information on system setup and problem severity levels was disregarded. When a transformer fails while it is operating, there are a variety of causes and corresponding warning indicators. Diagnosing a problem entail looking at the data, determining how serious the problem is, and ultimately pinpointing where in the system the problem lies [1]. Today, more and more data reflecting real-time condition information of transformer and an increasing number of information sources are accessible thanks to the implementation of novel detection techniques including on-line surveillance and live detecting of transformers. However, on just one hand, diagnostic procedures are defined by a single kind of characteristic data, such as ratio of gas pressure or contents, which leads to poor anti-interference capacity and simple to generate mistakes. However, diagnostic procedures may exhibit varying features, and analysis and use of such data may be hindered by a lack of coordination and complete analysis [2][3][4]. The present degree of information use is quite low. Experimental modal analysis may identify damage by observing shifts in a structure's inherent frequencies, damping ratios, and mode shapes (EMA). To pinpoint the damage's origin and extent, it compares the observed modal characteristics to either a baseline or fictional value and looks for deviations. However, erroneous conclusions may be reached as a result of processing such a big and complicated data set. In a damage detection system inspired by the Dempster-Shafer theory of evidence, a sort of multi-information fusion was used. When compared to more traditional single-parameter approaches, the built-in parameter's sensitivity and reliability were both significantly improved. The findings also demonstrated the broad applicability of the multi-information fusion approach to damage diagnosis. The rising incidence of fatal and seriously injured traffic accidents is a major contributor to these tragic outcomes. Having up-to-date information on incidents is crucial. At the moment, systems designed to identify accidents are either improving their detection accuracy or trying to make accidents more catastrophic. Knowing the specifics of an event helps emergency medical services (EMS) provide the best care possible to those affected [5][6]. In order to properly treat patient-specific pathology early on, longitudinal imaging may record both fixed anatomical features and the dynamic changes that accompany illness development. However, longitudinal data is seldom used in traditional methods of diabetic retinopathy (DR) detection. Automatic depression identification has come a long way in recent years, and a lot of that progress may be attributed to modality fusing and deep learning techniques. Multi-modal techniques, on the one hand, add a great deal of complexity to the data collecting phase, while deep learning methods, on the other, are notoriously mysterious, which undermines their trust. In this study, we present a multi-task BLSTM model that can process text using pretrained word embeddings. With a state-of-the-art F1 score of 0.87, our technique outperforms prior multi-modal research and provides data for the existence of depression as well as a projected severity score [7]. Moreover, we get the best RMSE among all existing text-based methods. Finally, we use a per-time-step attention mechanism to analyse the statements that contributed most to forecasting the sad state. Words like "um" and "uh" and other paralinguistic information are really strong predictors of sadness in our approach. For the first time, the fact that conversational fillers may notify a deep-learning model to the possibility of depression has been exposed. Purpose Improve diabetic retinopathy symptom awareness and disease severity categorization from fundus pictures using a hybrid architecture (DR) [8]. Techniques For this study, we utilised 26,699 fundus pictures of 17,834 people with diabetes from three hospitals in Taiwan, collected between 2007 and 2018, to categorise DR severity. A total of 37 ophthalmologists served as the actual truth for the pictures' lesion diagnosis and severity rating. Two different fusion architectures were proposed for deep learning: one that combines lesion detection and categorization models sequentially to mimic the choice process of ophthalmologists (late fusion) and another that merges lesion and intensity classification techniques in parallel (postprocessing fusion). The efficiency of the architecture was tested using Messidor-2 and 1748 pictures. Classification accuracy, the weighted statistic, and the area underneath the receiver operating characteristics curve were the major indicators of performance (AUC). Results With an accuracy and weighted of 84.29% and 84.01%, correspondingly, for five-class DR grading, a hybrid architecture performed well when used to hospital data. This method also improved upon the accuracy of traditional classification algorithms in identifying photos of early DR [9]. For referral DR detection, the Messidor-2 model scored an AUC of 97.09%, whereas state-of-the-art techniques trained on a bigger database produced AUCs of 85% to 99%. Conclusions In order to make our hybrid designs more durable and trustworthy for widespread use, we enhanced them by extracting features from DR pictures and boosting the effectiveness of DR grading. Functional Relevance for Translation The suggested fusion designs have the potential to provide for more rapid and accurate identification of several DR disorders than is currently possible via manual clinical practice alone [10].

II. Related Work

Using machine learning models like Random Forest, Gradient Boosting, and Xtreme Multilayer Perceptron trained with patient data and deep learning models like DenseNet201 and InceptionResNetV2 received training with knee x-ray images, this study aims to enhance the identification of Knee Osteoarthritis at all levels based on the Kellgren-Lawrence scale. The cumulative predictive power of these models is used as the basis for the final classification decision, which is made via a late fusion technique. Both machine learning and deep learning models obtained superior performance as measured by Precision, Recall, and F1-score, however the ROC curves for deep learning models revealed a greater level of efficiency. In addition, the primary characteristics that drive the illness were found using patient data in models of machine learning [11]. To identify inter revolution short circuits in inductors, this work introduces an approach based on fusion, which is applied to signatures derived from external stray flux. The belief function framework is used by this method to describe and combine data collected from sensors located all over the system in order to identify any potential short circuits. Researchers investigate how different sensor placements around a machine affect their ability to spot problems. The data collected from the machine's ambient magnetic field is fused with other sources to create a novel diagnostic procedure with the benefit of being non-invasive. The diagnostic method relies on data collected by six external flow sensors mounted on a belt around the machine. These fingerprints are produced by experiments with a rewound induction machine, which may generate varying degrees of inter-turn short circuit defects [12]. The purpose of this work is to investigate the identification and severity of thermal defects in electrical and mechanical automation equipment. Before proposing an MSIF method grounded on the D-S evidential theory, this work investigated the heterogeneity Multi-Sensor Information Fusion (MSIF) challenge presented by sensors located in crucial areas of electrical and mechanical automation equipment. The impact of thermal problems on the various equipment components was then assessed, with the report offering supporting evidence for the placement of sensors in strategic locations. Finally, testing findings demonstrated the efficiency of the proposed approach, and the thermal defect detection results were reached [13]. Using hierarchical labelling, the authors of this research present a unique data-driven technique for fusing system structural knowledge and describing severity levels inside a single learning framework. We derive the Large Margin Information Fusion (LMIF) technique and create a streaming-data-specific online learning system. The suggested approach is applied to the FDD of a 90-ton centrifugal water-cooled chiller, in accordance with ASHRAE Research Project 1043 (RP-1043). The benefits of merging past knowledge of fault dependency data into the learning process are supported by experimental data showing that LMIF can significantly enhance the FDD performance and detect the faulty risk level with high accuracy [14]. So as to increase the accuracy of a diagnostic, this study presents the information energy technology and proposes a new criterion that should be efficiently integrated and logically stratified [15].

III. Proposed Work

Several different techniques exist for accurately classifying and predicting medical conditions. Data mining for medical practise faces a number of obstacles and problems.

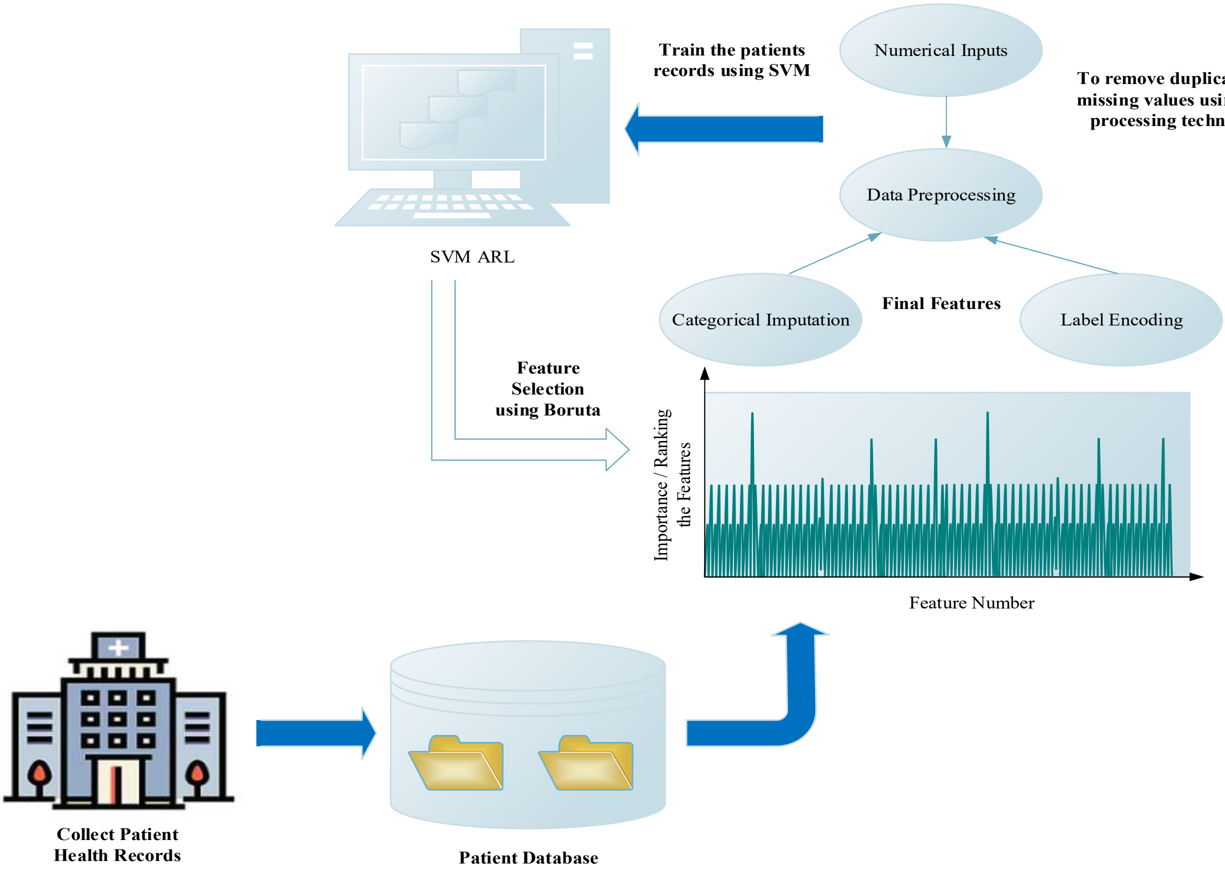

Fig. 1. Suggested work flow

A. *Data Pre-processing*

Date, Numeric data, and String, types are represented in the dataset's columns. In addition to continuous variables, our dataset also contains categorical ones. The categorical variables were label-encoded since the ML model expects all input data to be in numerical form. This procedure gives a numeric value to each distinct column category. There are a number of blanks in the data set that, when used as input, will result in a failure. Thus, we use "NA" to represent missing data. The "death" and "recov" columns of certain patient collected data are both blank; these records were extracted from the occupy a central place and added to the testing dataset, while the remainder records were added to the train dataset. Date-formatted columns also exist in the dataset. Instead of directly using the data columns, model. this model has been employed. The relevant (hosp vis—sym on) value has been entered into a new column. From this, we may calculate how long it was until the patient went to the hospital after first experiencing symptoms.

B. *Machine Learning Classifiers*

Logistic Regression. For binary and multiclass issues, logistic regression is a popular statistical classification approach. Logistic functions are used to forecast the likelihood of a class label. Hypothesis in its operational version is

$$\mathrm{Y} = \mathrm{C}^{\mathrm{T}}(\mathrm{X}) \tag{1}$$

where C is a list of linear regression and X is a list of attributes.

$$\mathrm{C} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \cdots \\ \beta_n \end{bmatrix}, \mathrm{X} = \begin{bmatrix} \mathrm{X}_1 \\ \mathrm{X}_2 \\ \mathrm{X}_3 \\ \cdots \\ \mathrm{X}_n \end{bmatrix} \tag{2}$$

where β_i projected weights, or regression estimators, for the characteristics of interest in the data, and β_0 symbolises the point when an equation begins to slope down.

$$(x) = Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n \tag{3}$$

Our study's logistic regression technique uses a total of 25 features from the dataset, therefore it is based on the dataset's characteristics.

$$\mathrm{h(x)} = (\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n) \tag{4}$$

If the value of is less than or equal to, the model will classify the record as a survivor or a fatality

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n \geq 0 \tag{5}$$

The maximum-likelihood ratio notion is used to find the best possible regression estimator. The attribute-to-class label mapping is performed using a sigmoid function (logistic function).

The following equations describe the sigmoid equation in its functional form:

$$S(g) = \frac{1}{(1+e^{-y})}$$
$$S(g) = \frac{1}{\left(1+e^{-C^{T}(X)}\right)} \quad (6)$$

where e is a numeric constant Euler's number. To prevent model overfitting, LR employs a regularisation parameter. Hyperoptimized parameters for the logistic regression were found using a grid search optimization.

Moreover, feature selection may be accomplished with the help of a random forest. It divides the data into a training set and a test set with the help of bootstrapping data sampling. Every bootstrap uses a different set of trees that are generated repeatedly by the model. The final forecast is calculated by averaging the votes from each category. It's the sum of all the branches of the many decision trees. An example of a classification method that uses a hierarchical structure is the decision tree. information gain, gain ratio and Entropy the Gini-index are all used to determine which node should be the decision maker. Both data augmentation and entropy were used in our analysis, as shown by the following formula as:

$$E(Y) = \sum_{i=1}^{n} - p_i \log_2 p_i$$
$$E(X, Y) = \sum_{n \in X} P(n) E(n) \quad (7)$$

where $E(Y)$ represents the entropy of the target, while Entropy (X, Y) is the entropy of the attributes with the target, in which $X = \{x_1, x_2, \ldots, x_n\}$ is the list of characteristics that describe the data. As a result, the attribute that yields the most useful data will be the "root" attribute:

$$\text{Information\_Gain} = E(Y) - E(X, Y) \quad (8)$$

It takes the predictions of many trees and mixes them using a set of random vectors denoted by _T. None of the vectors you choose will affect the vectors you chose before. The resulting forest of trees may be represented as $h(x)$. The following diagram illustrates the decision tree's tendency to mistake in its overall conclusions:

$$GE = P_{X,Y}\,(\text{margin\_fuc}(X, Y) < 0) \quad (9)$$

where $P_{X,Y}$ quantifies how likely it is that a given collection of characteristics corresponds to the Y category.

The XGB algorithm is a classification and regression method that uses an ensemble of models. The gradient boosting technique in its regularised version. Unfortunately, model overfitting occurs sometimes in the gradient boosting approach owing to the asymmetry of the data. Although overfitting is still possible, the XGB algorithm's regularisation parameter helps mitigate this. Comparable to randomized forest, XGB is a tree-based ensemble learning method. The goal of the boosting data resampling technique is to improve the accuracy of the model by reducing the misclassification error. It's an iterative method. In the following iteration, the model was trained using the data that had previously been used for unsuccessful predictions. The method will be repeated until the model reaches a state of optimality.

The regularisation parameter lowers the model's variance by giving misclassified occurrences a larger share of the total weight. Addition of mass ameliorates model underfitting. To prevent overfitting from causing a significant increase in the misclassification rate, penalty regularisation was used to mitigate the model's inherent bias. The XGB methodology is the result of adjusting a number of variables simultaneously. Better model performance may be achieved by finding the sweet spot for parameter values. The grid search method was used to find optimal values for the parameters.

Rule extracting from such a learned SVM (SVM-Rule) process is crucial for data mining as well as the discovery of knowledge because to the SVM's high accuracy in regression and classification. Nevertheless, in fact, the rules derived using SVM-Rule are less intelligible than we expected, due to the presence of a large number of obscure numeric factors (i.e., support vectors) in those rules.

While the decision tree's obtained rules might not be as precise as SVM rules, they are simple to understand because each rule corresponds to a single decision path that can be followed in the tree structure. Using the coordinates from SVMs aggregate in the tree, as well as the tree structure as a result of a decision tree rule, rule extracting using SVM trees (SVMT-Rule) is able to do rule extract over a tree structure of SVM. The support vector rule maintains the excellent classification accuracy, while the decision-tree rule improves the rule's understandability. And as the SVMT Rule aggregates groups of SVMs, it may conduct a very accurate classification on datasets with severe, even overpowering, class-imbalanced data distribution.

IV. Results & Discussion

Numpy, Pandas, SciPy, Scikit Learn, Matplotlib and Datetime, are some of the required packages and libraries for this endeavour. The project's execution has been placing on Google Colab, using the CPU real-time. Google Colab has a CPU with the following specifications: model 79, CPU Family 6 (Intel Xeon), model name 79, 2.20 GHz (Cache: 56,320 KB), and model name 6 (CPU Family 6). Google Drive is being utilised for storage.

Common metrics including precision, sensitivity, specificity, accuracy, and the F-score were used to assess the model's efficacy. The classifiers were compared using a number of different metrics, including area under the curve and aspects involved characteristic (ROC). It is one of the most used methods for evaluating the diagnostic test's sensitivity (response rate) and specificity (false-positive rate).

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (10)$$

where proper classification of test data is a reflection of the model's accuracy.

$$\text{Sensitivity} = \frac{TP}{TP+FN} \quad (11)$$

Sensitivity measures how well a classification system can anticipate positive labels. The positive predicted value (PPV) or true positive rate (TPR) is another name for this statistic (PPV).

$$\text{Specificity} = \frac{TN}{TN+FP} \quad (12)$$

The sensitivity of a classification system is measured by its ability to accurately forecast the frequency with which negative class labels will be assigned.

$$\text{F}-\text{score} = \frac{2\times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \quad (13)$$

F-score is the average of how well something was remembered and how accurate it was. The performance was given in fig 2 to fig 6 respectively and the table 1 gives the performance analysis.

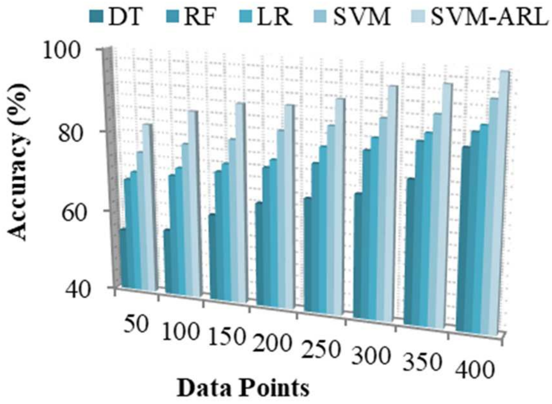

Fig. 2. Accuracy vs. Data Points

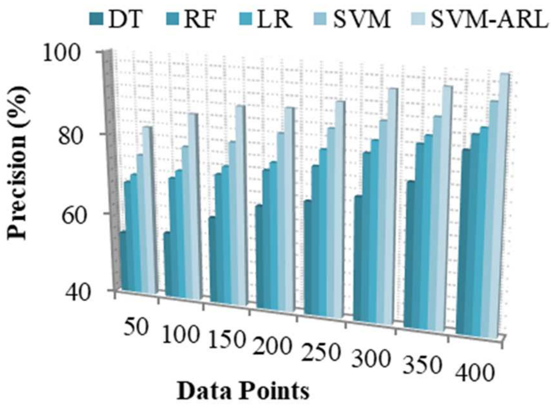

Fig. 3. Precision vs. Data Points

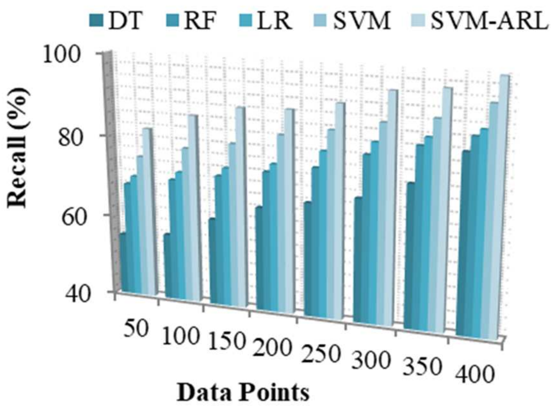

Fig. 4. Recall vs. Data Points

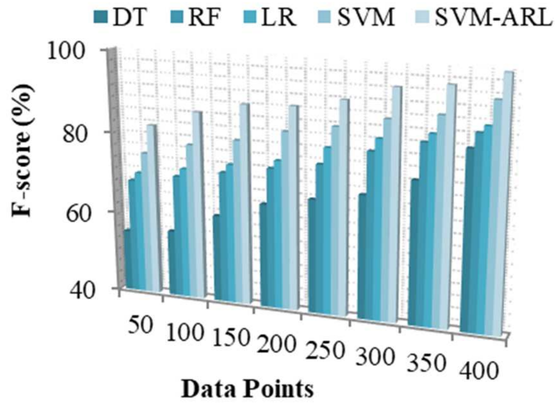

Fig. 5. F-score vs. Data Points

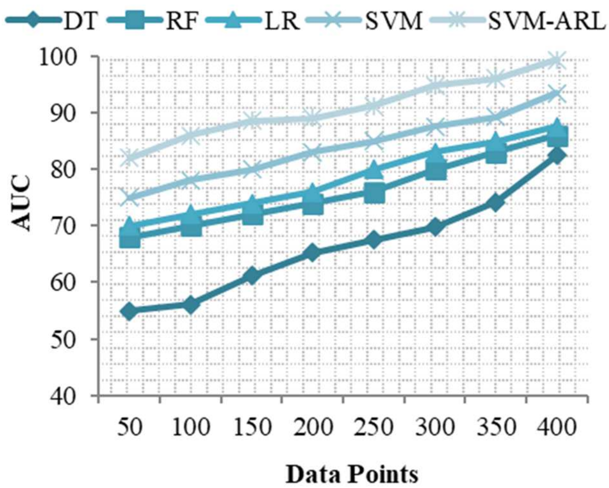

Fig. 6. AUC Curve vs. Data Points

Nevertheless, as we were particularly interested in healthcare and the practical application of machine learning to monitoring heart rate, we thought it was important to assess how well existing algorithms predicted anomalies for a real-world data set of heart rates. To determine how often the algorithms miss outliers and how much the preceding training phase on simulated data affected anomaly identification accuracy on actual heart rate data, we resorted to visualisations. Prediction techniques with poor sensitivity will not perform well in their main purpose (e.g., the detecting of anomalies), and prediction models with low sensitivity will result in an elevated false alarm rate, therefore it is crucial to bear in mind the relevance of both while using these algorithms. By keeping both of these systems performance metrics at high levels when these anomaly - based models are deployed in the real world, we can assure that both medical professionals and patients can have faith in the system's predictions.

TABLE I. EVALUATION ANALYSIS

Methods	Accuracy	Precision	Recall	F1 Score	AUC
DT	0.78	0.81	0.74	0.77	0.76
RF	0.81	0.87	0.73	0.79	0.81
LR	0.65	0.64	0.71	0.67	0.65
SVM	0.78	0.84	0.69	0.76	0.80
SVM ARL	0.84	0.86	0.81	0.83	0.84

V. CONCLUSION

A difficult but crucial endeavour is the automated identification of irregularities in physiological parameters, such as monitoring of heart rates. The inherent unpredictability of human physiological data does contribute to the challenge of anomaly detection in health data. Consider the case of a person whose current heart rate is flagged as abnormal by an anomaly detection system but who, in fact, is just starting a very strenuous exercise routine. This study analyses the performance of five algorithms based on machine learning in identifying abnormalities in heart rate data by training them on two synthetic datasets. We used the MIT-BIH database to get heart rate data from a single patient and test five different models.

REFERENCES

[1] Tseng, V.S., Chen, C., Liang, C., Tai, M., Liu, J., Wu, P.Y., Deng, M., Lee, Y., Huang, T., & Chen, Y. (2020). Leveraging Multimodal Deep Learning Architecture with Retina Lesion Information to Detect Diabetic Retinopathy. Translational Vision Science & Technology, 9.

[2] Gao, Y., Liang, Z., Wang, B., Wu, Y., & Wu, P. (2018). Wetland Change Detection Using Cross-Fused-Based and Normalized Difference Index Analysis on Multitemporal Landsat 8 OLI. J. Sensors, 2018, 8130470:1-8130470:8.

[3] Cessna, J., Alonzo, M., Foster, A.C., & Cook, B.D. (2021). Mapping Boreal Forest Spruce Beetle Health Status at the Individual Crown Scale Using Fused Spectral and Structural Data. Forests.

[4] Datla, R., Chalavadi, V., & Mohan, C. K. (2022, March). A framework to derive geospatial attributes for aircraft type recognition in large-scale remote sensing images. In Fourteenth International Conference on Machine Vision (ICMV 2021) (Vol. 12084, pp. 172-179). SPIE.

[5] Dousa, A., & Mohamed, S.A. (2016). Data fusion of video camera and Laser Range Finder (LRF) for obstacle avoidance.

[6] Bocklet, T., Steidl, S., Nöth, E., & Skodda, S. (2013). Automatic evaluation of parkinson's speech - acoustic, prosodic and voice related cues. Interspeech.

[7] Nallagonda, S., Godugu, K.K., & Ranjeeth, M. (2020). Energy-Efficiency Analysis of Cognitive Radio Network with Improved Energy Detectors and SC Diversity over Nakagami-q Fading Environment. 2020 IEEE International Symposium on Sustainable Energy, Signal Processing and Cyber Security (iSSSC), 1-6.

[8] Ramesh, G. P., Aravind, C. V., Rajparthiban, R., & Soysa, N. (2014). Body area network through wireless technology. Int. J. Comput. Sci. Eng. Commun, 2(1), 129-134.

[9] Lu, P., Huo, C., Duan, W., Ai, J., Jin, H., & Jin, L. (2019). Information Fusion and Image Processing Based Arc Detection and Localization in Pantograph-Catenary Systems. 2019 22th International Conference on Information Fusion (FUSION), 1-8.

[10] Saucedo-Dorantes, J.J., Zamudio-Ramírez, I., Cureño-Osornio, J., Osornio-Rios, R.A., & Antonino-Daviu, J.A. (2021). Condition Monitoring Method for the Detection of Fault Graduality in Outer Race Bearing Based on Vibration-Current Fusion, Statistical Features and Neural Network. Applied Sciences.

[11] Abasolo, K.H. (2021). Detection of Knee Osteoarthritis Severity using a Fusion of Machine and Deep Learning models.

[12] Irhoumah, M., Mercier, D., Pusca, R., Lefevre, E., & Romary, R. (2017). Information fusion of external flux sensors for detection of inter-turn short circuit faults in induction machines. IECON 2017 - 43rd Annual Conference of the IEEE Industrial Electronics Society, 8076-8081.

[13] Perveen, N., Singh, D., & Mohan, C. K. (2016, December). Spontaneous facial expression recognition: A part based approach. In 2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA) (pp. 819-824). IEEE.

[14] Li, D., Zhou, Y., Hu, G., & Spanos, C.J. (2016). Fusing system configuration information for building cooling plant Fault Detection and severity level identification. 2016 IEEE International Conference on Automation Science and Engineering (CASE), 1319-1325.

[15] Wang, X., Wang, S., & Zhu, W. (2019). Multilevel Information Fusion DGA Diagnosis Method for Power Transformer. 2019 IEEE 3rd Information Technology, Networking, Electronic and Automation Control Conference (ITNEC), 770-773.