THE RECENT RESEARCH ON MACHINE LEARNING ALGORITHMS

Yadav Sangeeta Ramchandra[1] and Sanjay Singh Bhadoria[2]
[1]Department of Computer Science & Application, Dr A.P.J Abdul Kalam University, Indore (M.P), India
ghodke.sangeeta@gmail.com
[2] Department of Computer Science & Application, Dr A.P.J Abdul Kalam University, Indore (M.P), India

ABSTRACT
Machine learning (ML) has witnessed unprecedented growth in recent years, driven by advancements in algorithms, increased computing power, and the availability of vast datasets. This paper provides a comprehensive review of the recent research on machine learning algorithms on data analytics, covering a wide range of applications, techniques, and challenges. The discussion encompasses both traditional and novel approaches, shedding light on their applications, strengths, and limitations. We also explore emerging trends, ethical considerations, and potential directions for future research in the dynamic field of machine learning.

Keywords- Machine Learning, Big Data, Artificial Intelligence, Neural Network, Data Analytics

INTRODUCTION
Integration of online data systems is one of the most rapidly increasing and rising topics in computer-related applications. Entities can speak with one other thanks to recent advancements in communication technologies. Entities may communicate, listen to, and respond to one another as well as to their surroundings. It is projected that between 25 and 50 billion internet devices would be deployed for varied application usage by 2020 [1]. In the Internet data, actuators and sensors are installed in the external world to perceive parameters. These gadgets communicate with one another via communication networks. At the collection centres, the same network is utilised to provide the measured parameters in the form of raw data. Data science gives a new and substantial possibility to make online data applications more intelligent, since internet has become one of the most important sources of raw data/new data.

The framework of internet computation is a key feature. According to [2] based on the location of the internet data processing, it may be classified as follows:

1. Edge Computing: Processing performed on internet devices.
2. Data from data sources is processed using fog computing.
3. Distributed computing is utilized for data processing.

Wide range of applications, including healthcare, industrial units, smart transportation, smart parking, smart grid, entertainment sector [3], and remote sensing, to mention a few, areas utilising online data systems to generate and store data on cloud servers. Policymakers, industrialists, and physicians, among others, utilises analytical methods to analyse this data and obtain meaningful insights. This insight is utilised to improve existing services and to supplement those supplied to its stakeholders [4]. This knowledge may be accessible at any time and from any location by utilising any communication device.

Machine learning (ML) is the application of artificial intelligence (AI) to a system in which the system learns from previous experience without being explicitly programmed. Its primary goal is to learn without human intervention or support and to alter the behaviour of the systems [5]. For learning, systems employ a variety of models. These models may be designed to learn online by utilising data streams generated by ubiquitous devices, which are referred to as stream learning or online learning. Models may also learn from previous data, which is known as offline learning. A learning system can be classified into one of three types: supervised learning, unsupervised learning, or reinforcement learning [6].

There is a rising need and desire in the research community for a generic approach that can be employed across a varied range of applications for different classification issues in order to analyse complicated current data gathered from multiple sources [7]. Unfortunately, the majority of machine learning (ML) research is focused on a well-defined and extremely specialised topic in a certain application domain. [8] To answer the problem of complicated aggregates of data, researchers must focus their efforts on designing and developing a system that can handle a wide range of problems and data kinds. [9] The new system must compete with state-of-the-art approaches for solving certain challenges. There are four objectives that are to be fulfilled. They are 1) Developing the machine learning algorithms that can computationally scale to Big data 2) Designing algorithms that do not require large amounts of labeled data, 3) Designing a resource-efficient machine learning methods, and 4) developing a privacy preservation techniques for various applications.

LITERATURE REVIEW

Engineering asset resilience management critically depends on the ability to detect events that may prevent assets from safely and dependably delivering their intended function. While the value of the data asset itself is increasingly acknowledged, collected data are often left poorly exploited, failing to take appropriate advantage of their potential for enhancing asset management performance [9]. Part of the challenge lies with the difficulty in understanding when any observable change in the data corresponds to events of interest, which in turn must require intervention actions. Directing the attention on significant events remains a challenging problem in asset monitoring. Terms such as outlier detection [10], novelty detection [11], and anomaly detection [12], are all employed in this context and are relevant to change detection when monitoring engineering assets [13]. However, collected data are typically not linked with validated event cases, making it hard to apply any supervised type of learning from data, making unsupervised [14] or semi-supervised types of learning more applicable in practice [15]. Most employed algorithmic approaches still require calibration and adjustment, as the dynamic characteristics of streaming data greatly vary across domains. Clearly, intrusion detection systems are exceeding barriers, which pose concerns about the safety of vital digital products [16]. ML, NLP, computer vision, and any more fields have emerged as a result of the development of AI. [17]. Many organizations still have limited resources in data management and/or computational power for big data analytics purposes [18]. The objective of big data analytics is to use available resources efficiently and predictively. The cloud provides a dynamic environment in which all these components are present, thus creating a readymade, economically efficient solution for such problems [19] [20].

Vulnerabilities in IoT devices create a huge number of opportunities for cyber security threats and other types of crimes, particularly among the networked devices that are already commonplace in homes detect cyber threats in real-time. [21] The constant rise in a variety of cyber threats and viruses clearly shows how inadequate the current defenses are for protecting computer networks and resources [22]. When compared to currently use non-machine learning-based models, our engine learning combined project reproduction outcomes demonstrate its movement resilience [23] operates effectively with regard to the failing point's restoration as well as business cost savings [24]. While maximum performance is an efficient wonder in industrial, research, and imitation technical domains, resilience is a more fundamental and accepted occurrence in the majority of trade, medicinal, and real-time areas [25]. The investigation reveals that Big Data, spatiotemporal thoughts, and diverse application contexts motivate the progression of cloud technology and related technologies with new standards [26], Big Data and cloud computing enable scientific findings and developing applications and inherent spatial-temporal fundamentals of Big Data and geospatial fields of science can provide a source for discovering technical and significant remedies for Big Data [27].

SUPERVISED LEARNING ALGORITHMS

This section explores recent developments in supervised learning algorithms, including:

Deep Learning Models: Review of state-of-the-art deep neural networks such as convolutional neural networks (CNNs), recurrent neural networks (RNNs), and transformer models. Transfer learning and pre-trained models for various tasks.

Ensemble Methods: Boosting and bagging techniques are in focus of recent advancements. Applications and improvements are in random forests and gradient boosting.

Kernel Methods: Advances in support vector machines (SVM) and kernelized models. It is in non-linear classification and regression.

Unsupervised Learning and Clustering: Recent research is based on unsupervised learning, dimensionality & clustering reduction techniques:

Variational Autoencoders (VAEs): Generative models and unsupervised representation learning gives Density-Based Clustering having advancements in density clustering algorithms and Embedding Techniques with word and graph embeddings are two types gives advances in preserving semantic relationships in high dimensional data.

Reinforcement Learning: Recent progress in reinforcement learning algorithms gives Deep Reinforcement Learning giving policy gradients and value-based methods and Multi-Agent Reinforcement Learning giving coordination and communication among agents

Ethical Considerations and Fairness: Addressing the ethical implications of machine learning, it explores recent research on fairness, accountability, and transparency in ML algorithms like Bias and Fairness for techniques for mitigating bias in algorithm with fairness aware machine learning & algorithmic fairness and Explainability giving interpretable machine learning models and model-agnostic explanations.

MACHINE LEARNING & DATA ANALYTICS

In all, 33 data sets from the UCI library were used [28]. These data sets were obtained utilising sensors. They correspond to various IoT domains, are of various sizes, and properties. To have a more diverse and generic data collection, just a few data sets are from non-IoT domains are taken. The binary and multiclass datasets utilised in this investigation are summarised in table 1. Three validation techniques are used to assess algorithm performance: hold out validation (HoV), 10-fold cross validation (10FCV), and 5*2 cross validation (5*2CV). These validation procedures regulate aspects like as training set selection and test set selection, which have a significant impact on MLA performance. These validation methods are neutral; however they have a lot of volatility. [29].

On all datasets, algorithms are run using each validation approach, and assessment is done using the most generally used classification job evaluation measure, "accuracy." Each validation approach is carried out in such a way that a total of 30 result samples are obtained. The final value for "accuracy" for each dataset is then calculated using an independent algorithm. Most critical parameter in security, smart lighting, smart transportation, and seismic hazard applications is accuracy. [30] One of the most crucial aspects of any research is statistical validation of outcomes. When all three requirements are met, a) independence, b) homoscedasticity, and c) normality, parametric tests are utilised [31]. Independency is not achieved when KFCV and 5*2CV are utilised, according to the authors of [32].

Table 2 shows accuracy of methods on specific datasets and accompanying rankings for binary class data using HOV. Similarly, average rankings are generated for all binary class data and multiclass data using 5*2CV and 10FCV.

ENSEMBLES OF DECISION TREE FOR ONLINE DATA ANALYTICS

Building a multi classifier system or having an ensemble of classifiers is a good and efficient technique to handle complicated data. It essentially mixes the hypotheses of many or disparate classifiers in order to generate a more accurate approximation of the genuine hypothesis.

Table 1: Data Set

Sr. No.	Name	Attributes	Instances
Binary class			
1	Extra Sensory-B	277	2686
2	EEG Signal	15	8123
3	Churn Modelling	13	10000
4	Card Mfg	16	690
5	Dota2 Games	116	102944
6	Electric Grid	14	10000
7	Liver Patient	10	583
8	Transitions Irish	5	500
9	Watch Sensors	13	7386
10	OC 01Driving	14	7392
11	OC 01 Road	14	7392
12	PC3	37	1563
13	PC4	37	1458
14	Diabetes	8	768
15	Stars Pulsar	9	17900
16	Bumps-Seismic	19	2584
17	Sonar	60	208
18	Hand Gestures	65	5811
19	Power System -B	99	4966
20	Machine Sensors	75	10616
Multiclass			
1	Air Quality sensor	16	9358
2	Cardiotocography	41	2126
3	Detection of Falls	6	16400
4	White Wine	11	4898
5	EMG	5	30000
6	Recog. Movement	563	2948
7	Abalone Original	8	4177
8	Avaliacao	24	30697
9	DC Nuclear	80	1081
10	Prediction Energy	29	19736
11	Big Sensors	516	6373
12	OC 01-Traffic	14	7392
13	Sky-Server	18	10000

OC: Opel Corsa, DC: Data Cortex ***(Survey)***

Researchers have been successful in establishing the benefits of ensemble learning (EL) for learning tasks due to the resilience and efficiency of ensemble approaches. Information retrieval, image classification, financial domain, sentiment analysis, natural language processing, online dynamic security assessment, and medical domain have all demonstrated the effectiveness of EL. Not only did EL do well in multiclass classification, but it also did well in multi-label classification. [33]. For information extraction, early efforts created learning strategies such as bagging, boosting, and stacking. Nature allows for both homogeneous and diverse ensembles. In big data, ELs are also employed as pre-processing techniques to eliminate noisy instances from data before converting it to smart data [34]. Researchers have also looked studying how feature selection approaches and different base classifiers affect EL.

For human activity recognition in a smart home context, the authors of [35] employed classifier chain, a multi-label classification technique. For the classifier chain technique, they compared bernoulli nave Bayes, decision tree (DT), logistic regression (LR), and K-nearest neighbour (K-NN) as base classifiers. The decision tree as a basis classifier beat the other base classifiers, according to the results.

Online Learning: Implement algorithms that can learn incrementally as new data arrives, allowing for continuous updates without reprocessing the entire dataset. This is particularly useful for streaming or dynamic big data.

Feature Selection: Identify and use only the most relevant features to reduce computational complexity.

Dimensionality Reduction: Techniques like Principal Component Analysis (PCA) can be employed to reduce the dimensionality of the data, making it more manageable without significant loss of information.

Dimensionality reduction techniques aim to reduce the number of features or variables in a dataset while preserving its essential information. Two widely used methods for dimensionality reduction are Principal Component Analysis (PCA) and t-Distributed Stochastic Neighbor Embedding (t-SNE). Here, I'll provide a simplified mathematical model for PCA, which is a linear technique, and for t-SNE, which is a non-linear technique

Principal Component Analysis (PCA)

Input Data: Let X be the original dataset with *n* samples and *d* features

The data matrix X is of size *nxd*, where each row represents a sample and each column represents a feature

Centering the Data: Subtract the mean of each feature from the corresponding feature values:$X_C = X - \bar{X}$ is the mean vector

Covariance Matrix: Compute the covariance matrix

$$C = \frac{1}{n-1} X_C^T X_C$$

Eigenvalue Decomposition: Perform eigenvalue decomposition on C=VDV*T*, where V contains the eigenvectors and D is a diagonal matrix with eigenvalues

Selecting Principal Components: Sort the eigenvectors based on their corresponding eigenvalues in D and select the top *k* eigenvectors to form the projection matrix W

Dimensionality Reduction: Project the centered data onto the subspace defined by W: reduced=$X_{reduced}$=$X_c \cdot W$

For reducing class noise from huge data, the authors of [36] created and contrasted homogeneous ensemble (random forest as basis classifier) and heterogeneous ensemble (random forest, K-NN, and logistic regression as base classifiers) using edited closest neighbour (ENN). The nature of this huge data is binary. When compared to

heterogeneous ensemble and ENN, the homogeneous ensemble strategy performed better not just in terms of accuracy but also in terms of computation time.

The k-Nearest Neighbors (KNN) algorithm is a simple, yet effective, machine learning algorithm used for classification and regression tasks. It can be described mathematically as follows: Suppose you have a dataset with feature vectors X_1, X_2,….X_n and corresponding labels Y_1, Y_2,…..Y_n, where each X_i represents a data point in the feature space and Y_i is its associated label. Define a distance metric, typically Euclidean distance, denoted as $d(X_i - X_j)$, to measure the distance between two data points X_i and X_j. Select a positive integer (k) (number of neighbors) for the algorithm. For a new data point (X_{new}) for which you want to predict the label Y_{new}, find the k nearest neighbors of X_{new} from the training dataset based on the chosen distance metric. Let N_{Xnew} be the set of indices of the k nearest neighbors of (X_{new}). Then, the predicted label Y_{new} for classification is given by the majority class among the Y_i's corresponding to the neighbors:

$Y_{new} = \text{argma}X_y \; i \epsilon N_{Xnew} \; 1(Y_i - y)$

Where 1(condition) is the indicator function, equal to 1 if the condition is true and 0 otherwise.

Prediction for Regression For regression tasks, the predicted value Y_{new} for a new data point X_{new} is often the mean (or median) of the Y_i's corresponding to its k nearest neighbors.

$Y_{new} = \frac{1}{k} \; i \epsilon N_{Xnew} \; Y_i$

This is a basic mathematical representation of the KNN algorithm. The key aspects are the distance metric, the choice of k, and the way predictions are made based on the neighbors. Different distance metrics and variations of the algorithm exist depending on the specific requirements of the problem.

The authors of [37] employed random forest and very randomised trees to classify breast cancer (ET). The classification accuracy, specificity, and sensitivity of these ensemble models were assessed. Both have been observed to have obtained 100 percent outcomes. This is an issue of binary classification.

In the study [38], cost-sensitive ensemble algorithms AdaC1, AdaC2, and AdaC3 (DT as basis classifier) are compared with 19 other single classifiers and ensemble classifiers to tackle the production quality classification issue in the realm of smart manufacturing. For each defect category, these algorithms are used to categorise the set circumstances of the die-casting manufacturing process as faulty or normal. AdaC2 excelled on all performance metrics, including the F1-score, G–mean, and AUC. This is a problem of binary classification with a class imbalance.

A decision tree is a supervised machine learning algorithm used for both classification and regression tasks. The basic idea of a decision tree is to recursively split the data based on certain features to create a tree-like structure that represents a sequence of decisions. A decision tree consists of nodes, where each node corresponds to a test on a particular feature. The edges between nodes represent the possible outcomes of the test, and the leaves of the tree represent the class labels. At each internal node, a decision is made based on a feature, and the data is split into subsets. The splitting criterion depends on the type of decision tree. For classification trees, common criteria include Gini impurity, entropy, or misclassification error.

$\text{Gini}(D) = 1 - \sum_{i=1}^{c} P_i^2$

where D is the dataset, c is the number of classes, and p_i is the probability of belonging to class i.

$\text{Entropy}(D) = - \sum_{i=1}^{c} P_i \; \text{Log}_2 \; (pi)$

where D is the dataset, c is the number of classes, and p_i is the probability of belonging to class i.

The performance of decision tree ensembles was evaluated in a paper published in [39]. Random forest, very randomised tree, rotation forest, gradient boosted tree, and Adaboosted tree are all examples of bagged decision

trees. A heterogeneous ensemble model is built with equal parts random forest, severely randomised tree, and rotation forest. In contrast to other tree-based classification models, it is discovered that heterogeneous ensemble provides the best rank. There are nine multiclass datasets utilised, each of with its own size and characteristics.

In a work published in [40], researchers tested 20 supervised ensemble learning techniques, including boosting, bagging, random forest, rotation forest, Arc-X4, class switching, and its variations, using 19 binary data sets of varying sizes and characteristics from the UCI repository. They also used isometric regression to look at the influence of two base learners, the highly randomised tree (ET) and the classification and regression tree (CART), as well as the effect of calibration on the model. The rotating forest algorithm family, with or without calibration, beats all other ensemble approaches. They used the signal-to-noise ratio (SNR) filter approach to choose a hundred of the most important attributes.

Smart gas data was utilised for user profiling in a research done by [41]. Gas data analytics is used to estimate fuel poverty and low-income group quality of life. The study employed machine learning techniques to classify data into four separate groups, each reflecting four different tariff kinds. Adaboost, decision tree, decision forest, decision jungles built on decision forest, neural network, support vector machine (SVM), and Bayes point machine are some of the machine learning techniques used for tariff classification (BPM). It was discovered that Adaboost and decision jungle had the maximum area under the curve (AUC) for tariff categorization, scoring 51.9 percent.

The authors of [42] used 10 alternative base classifiers for sensor-based human activity detection to perform performance study on Adaboost, an EL model. Individual learning models are used to assess performance. SVM, KNN, nave Bayes, RF, CART, C4.5, REPTree, and LADTree were the basis models employed. SVM has been shown to have the best performance among individual classifiers. On all performance criteria, Adaboost performance has improved with all base classifiers. Accuracy, F1-score, kappa, and AUC were the metrics employed.

The authors of [43] created a logitboost an EL model based on additive logistic regression for classifying IoT devices in a smart home environment based on traffic flow. They have derived many characteristics from raw data. They added statistical values to 59 extracted elements from traffic flow data as additional features. They used a multiclass classification model to divide data into four types based on the coefficient of variation ratio of traffic received and delivered from the device. Using independent characteristics, 99.79 percent accuracy was achieved in classifying devices into one of the established classes.

Gap Analysis

We did our best to find studies that only used a generalised ensemble model using a decision tree as the basic classifier. We were unable to locate any. The majority of the articles are devoted to specific challenges from various disciplines. As a result, we conducted a literature review on challenges in a specific domain.

To the best of our knowledge and based on the literature, no one has compared EL models for binary class and multiclass data independently with decision trees as the base learner, especially on internet datasets from diverse domains. Previous research has shown that ensemble models with a decision tree (DT) as the basic classifier outperform single methods. This section of the paper focuses on the comparative examination of variations of decision tree (DT) ensemble and their influence on other performance metrics utilising feature reduction approaches that do not require any parameter or calibration adjustment.

BONFERRONI DUNN TEST

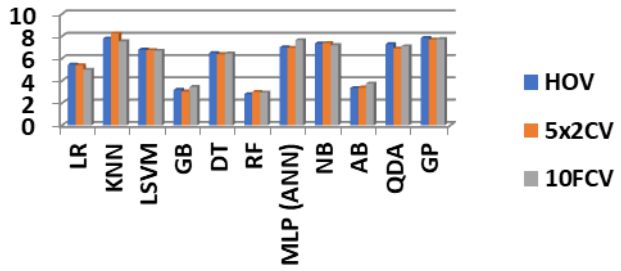
If the disparities in their average rankings across all datasets are more than or equal to the value of critical difference (CD) [44], the performance of two algorithms is considerably different (Dunn, 1961). Equation 3.4 is used to compute the critical difference (CD):

$$CD = q_{\alpha}\sqrt{\frac{K(K+1)}{6N}} \qquad (3.4)$$

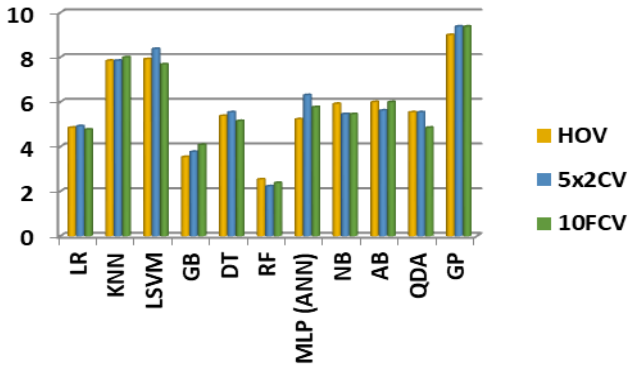
Where q_α= 2:773 is critical values (obtained from F-distribution table) for bonferroni dunn test which includes control algorithm. K is the number of algorithms and N is the number of datasets. The Bonferroni-Dunn graph for binary class data is shown in graph 3; with the bar height representing the average rank of the algorithm determined using Friedman ranking for the associated validation approach. The algorithm whose bar is higher than the threshold value (the sum of the average rank of the best performing algorithm, the control, and the crucial difference (CD)) performs poorly compared to the control algorithm (random forest). For each form of validation, the horizontal line in a graph represents a threshold value, which is equal to the average rank plus crucial difference (CD).

Graph 3 shows that, when using all cross validation approaches, the performance of decision Tree, SVM, MLP, QDA, NB, and KNN for binary datasets is inferior to that of RF. As depicted, the bonferroni-dunn graph for multiclass datasets in the same way.

We have not modified any of the parameters for any algorithms except for K-NN, where the values of K used are equal to 5 because k=5 has demonstrated the highest performance. This work aims to discover a generic model that may be extensively utilised in applications utilising internet sensors. For value of K, the trials are conducted from 1 to 5.



Graph 1: Bonferroni Dunn chart for Binary Class

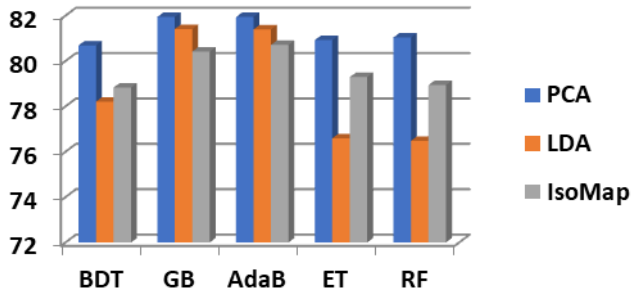


Graph 2: Bonferroni Dunn chart for Multiclass Data

Because it makes a careful final decision (either using the concept of voting or averaging or by increasing the weight of misclassified instances) based on decisions taken from its component weak base classifiers, ensemble learning models (ELs) have given better performance than single learning models for binary class and multiclass. It combines the advantages of several weak models. Random forest (RF) operates on both data and feature space at the same time. To accurately categorise the instance, it samples both the training sample space and the feature space. RF is the most extensively utilised ensemble approach in the health-care arena due to its simplicity and interpretability.

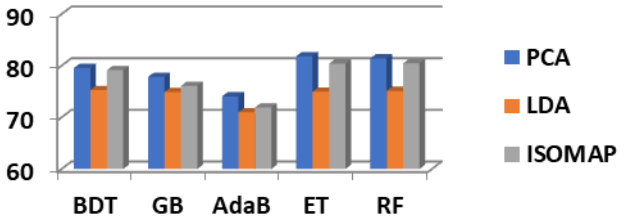
EXPERIMENTATION AND RESULTS

Google colab notebook, an online cloud-based tool, is used for experimentation. The code is developed in Python and employs the scikit-learn module. 5FCV is used for validation. Pre-processing is accomplished by removing the rows that contain missing values. The mean was used to replace out-of-range values. Values for the performance metric are averages across 10 binary data sets and 10 multiclass data sets for each feature reduction approach. Table 3 displays the averaged accuracy values derived by individual methods employing PCA, LDA, and IsoMap over 10 binary class datasets. The highest scores are in bold. Graph 3 depicts the appropriate graphs.



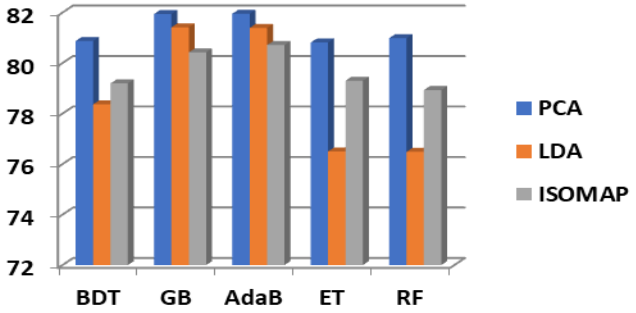
Graph 3: Averaged accuracy over Binary Data Sets using PCA, LDA and IsoMAP

From the graph depicted in graph 3 it is clear that Gradient Boost and Adaboost for PCA gives the best result for PCA over all others in case of binary data set accuracy.



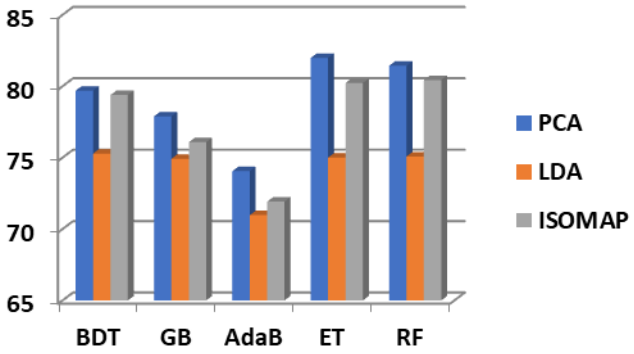
Graph 4: Averaged Accuracy over Multiclass Data Sets using PCA, LDA and ISOMAP

As per the graph 4 shown, it is clear that values of PCA are higher than ISOMAP and that it is higher than LDA. It can be also interpreted from the graph that Extreme Randamized Tree algorithm gives highest performance as compared to others. Even Randomized Tree is at par with it. Thus average accuracy of Extreme Randomized Tree algorithm for PCA is the best suited.



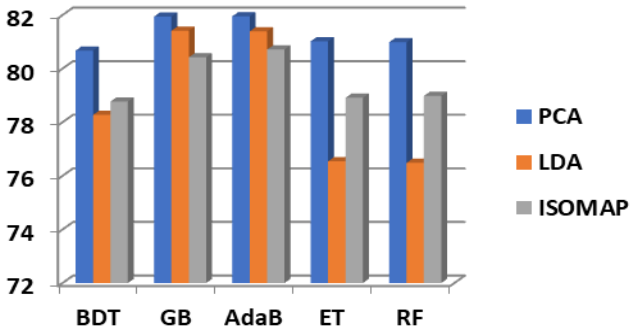
Graph 5: Averaged Precision over Binary Data Sets using PCA, LDA and ISOMAP

The graph depicted in graph 5 shows that Adaboost using PCA gives better results for precession for binary class of data set. The other Gradient boost is nearly same as the Adaboost for PCA. The other ones are not better than the two boosts.



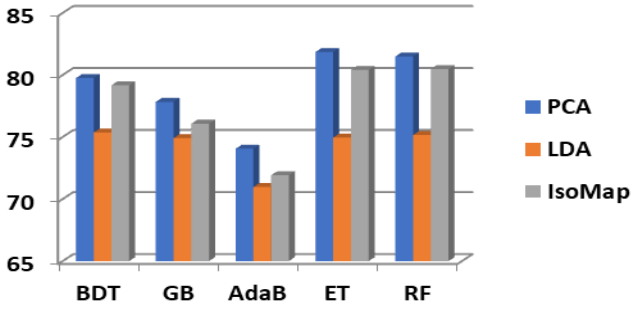
Graph 6: Averaged Precision over Multiclass Data Sets using PCA, LDA and ISOMAP

When precision for multiclass dataset was studied, it is observed as depicted in graph shown in graph 6 that, Extreme Randomized Tree algorithm for PCA is the best technique. Rest of the algorithms have not shown better result.



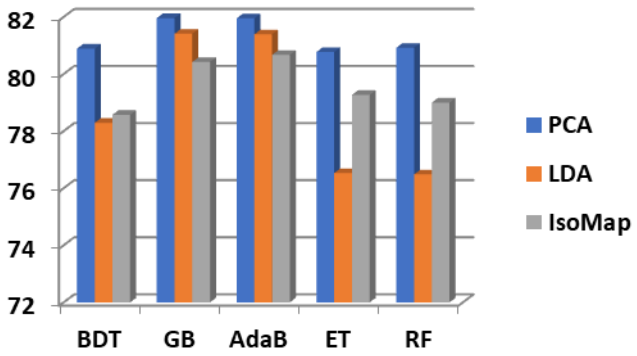
Graph 7: Averaged Recall over Binary Data Sets using PCA, LDA and ISOMAP

When recall of binary class was studied, it was observed that Adaboost and Gradient Boost have shown better performance applying PCA strategies. The rest of the techniques applied have not shown better result.



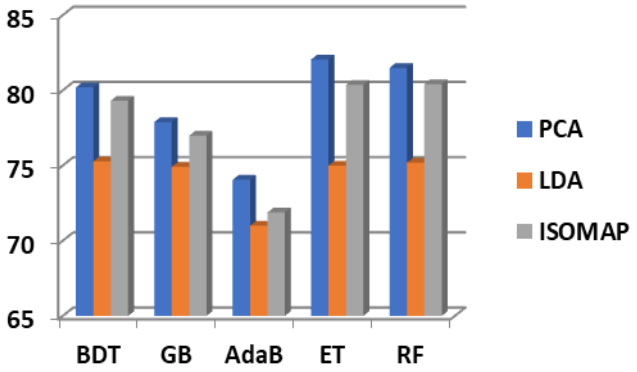
Graph 8: Averaged Recall over Multiclass Data Sets using PCA, LDA and ISOMAP

The recall of multiclass class data showed that, Extreme Randomized Tree and Randomized Tree algorithms using PCA strategies has given better performance. The rest of the techniques applied have not shown better result.



Graph 9: Averaged F1-Score over Binary Data Sets using PCA, LDA and ISOMAP

The average F1 Score for binary class has been observed as depicted in graph 9 and it can be interpreted that, Adaboost and Gradient boost of PCA has shown higher performance when compared with all others.



Graph 10: Averaged F1-Score over Multiclass Data Sets using PCA, LDA and ISOMAP

It is shown in the graph 10 that, Extreme Algorithm of PCA has been the highest performance entity amongst all others. This was calculated considering averaged F1 Score over Multi class data sets.

RESULTS

ACCURACY

Binary: The performance difference between PCA and LDA for all methods ranges from 2.5 to 4.5 percent, with the exception of gradient boost and Adaboost, which are both 0.5 percent. The difference between PCA and IsoMap is 1.5 to 2.0 percent. PCA, Gradient boost, and Adaboost have comparable performance and are the top scorers. They differ from other algorithms by about 1%.

Multiclass: The observed difference in performance for algorithms ranging from PCA to LDA is around 3% to 7%, which is substantial. The difference between PCA and IsoMap is between 1% and 2%. Using PCA, ET and RF fared better with comparable performance and are the top scorers. They differ from other algorithms by around 2.55 percent to 7.55 percent, which is much more.

PRECISION

Binary: The reported difference in performance for algorithms ranging from PCA to LDA ranges from 2.5 to 4.5, with the exception of gradient boost and Adaboost, which are both 0.5. The difference between PCA and IsoMap is between 1.5 and 2.5. Using PCA, gradient boost and Adaboost work similarly and get high scores. They differ from the others by around one.

Multiclass: The observed difference in performance for algorithms ranging from PCA to LDA is considerable and ranges from 3 to 7. Their ratio of PCA to IsoMap is 0.5 to 2. Using PCA, ET and RF outperformed with comparable performance and are high scorers. They differ from others in a range of around 3.0 to 7.0.

RECALL

Binary: The reported difference in performance for algorithms ranging from PCA to LDA ranges from 2.5 to 4.5, with the exception of gradient boost and Adaboost, which are both 0.5. The difference between PCA and IsoMap is between 1.5 and 2.0. Using PCA, gradient boost and Adaboost work similarly and get high scores. They differ from the others by around one.

Multiclass: The observed difference in algorithm performance between PCA and LDA is considerable and ranges from 3 to 7 roughly. Their ratio of PCA to IsoMap is 0.5 to 2. Using PCA, ET and RF outperformed with comparable performance and are high scorers. They differ from others in a range of around 3.0 to 7.0.

F1-SCORE

Binary: The reported difference in performance for algorithms ranging from PCA to LDA ranges from 2.5 to 4.5, with the exception of gradient boost and Adaboost, which are both 0.5. The difference between PCA and IsoMap varies between 1.5 and 2.0. Using PCA, gradient boost and Adaboost work singes from 3 to 7. Their ratio of PCA to IsoMap is 1.0 to 2. Using PCA, ET and RF outperformed with comparable performance and are high scorers. They differ from others by a factor of about 2.0 to 8.0.

Following experimenting and charting the findings on bonferroni-dunn charts for binary and multiclass data independently, the following final conclusions may be reached In the case of binary class data, random forest, gradient boosting, and Adaboost using decision tree as the basis classifier performed better on the accuracy scale than other classifiers utilised. Logistic regression has provided decent results but is close to the cut-off value. In the case of multiclass data, random forest and gradient boosting performed better in terms of accuracy. Logistic regression has also performed well, although it is closer to the threshold. It may also be deduced that not all algorithms perform similarly on all types of issues. They outperform other ML algorithms substantially.

As a result, it is reasonable to assume that ensemble models should be chosen over individual algorithms when no historical data distribution information is available. With these findings, we will investigate the impact of alternative ensembles of decision tree (DT) on the IoT data set. It is also vital to investigate how the feature reduction strategy affects other performance measures. After analysing the findings and comparing performance metric values, it is possible to infer that gradient boost and Adaboost should be favoured over others for binary classification with PCA or LDA. They regularly scored well in terms of accuracy, precision, recall, and F1-score.

Extremely randomised tree and Random forest techniques should be selected over other multiclass classification algorithms that employ PCA as a reduction approach. When compared to the others, ET and RF performed better. They regularly scored well in terms of accuracy, precision, recall, and F1-score.

TESTING OF DATA AND HYPOTHESIS

Algorithms function differently when dealing with binary and multiclass data.

For this hypothesis, it was required to analyse three different cross validation techniques used to assess the performance of the algorithm. The three techniques analysed were hold out validation (HOV), 5x2 cross validation (5x2CV) and 10 fold cross validation (10FCV). Anova test was applied to data separately of each of the techniques regarding average ranks for binary and multiclass data.

The result of analysis gave three different *P* values. For HOV analysis the *P* value calculated is 0.00157, for 5x2CV it is 0.000863 whereas for 10FCV it is 0.003823. If considered individually or collectively, the P values received are below 0.05 giving an indication that the hypothesis is accepted. Therefore it can be said that "Algorithms function differently when dealing with binary and multiclass data".

Suggested model for generalized learning for the online domain, outperforms, or is on par with state-of-the-art models.

As discussed above for hypothesis H2, data about average accuracy for ensemble models, gives meaning about average area under the curve (AUC) and data about average F1 score for ensemble model using PCA, LDA and ISOMAP. Here F1-score is the evaluation matrix that combines two matrices: Precision and Recall, into a single metric by taking their harmonic mean. It is also called as F-measure. Here accuracy, area under the curve (AUC) and F1 score are considered for analysis.

On analyzing the data, it is observed that calculated P value is very less i.e. 4.19853×10^{-78}. This is far below 0.05 that is sufficient for proving the hypothesis. Hence the hypothesis is accepted. Hence it can be said that "On most performance criteria, the suggested model for generalized learning for the online domain, outperforms, or is on par with state-of-the-art models"

CHALLENGES AND FUTURE DIRECTIONS

The paper discusses the current challenges in machine learning and proposing potential avenues for future research.

Scalability and Efficiency: Handling large-scale datasets and real-time processing developing resource-efficient algorithms.

Interdisciplinary Collaborations: Integrating domain knowledge into machine learning models giving collaborative research across disciplines for more effective solutions.

Security and Privacy: Robustness against adversarial attacks by privacy-preserving machine learning techniques.

Continual Learning and Lifelong Adaptation: Developing algorithms capable of continuous learning adapting to changing environments and emerging data patterns.

CONCLUSION

This work has demonstrated, as part of its first contribution, that the performance of ML algorithms when employed for generalised learning for diverse IoT applications is not identical. They differ in terms of accuracy, which is one of the most extensively used performance metrics in machine learning. They not only perform differently, but their performance differences are considerable. Non-parametric tests were used to statistically validate the experimental data. It is also found that when accuracy results were compared, all ensemble models (random forest, Adaboost, and gradient boost using decision tree as basis classifiers) performed better as a generic learner in the area of online data for binary class and multiclass data.

Our next contribution looked at various ensemble models that used a decision tree as the basic learner. It also considered linear and non-linear feature reduction approaches. Other performance parameters, including as precision, recall, and AUC, were also taken into account. According to the experimental data, gradient boost and Adaboost are the best options with PCA for accuracy, precision, recall, and F1-score. The highly randomised tree (ET) and random forest with PCA performed best on accuracy, precision, recall, and F1-score for multiclass data.

The final result of this research was the development of a unique technique dubbed "hybrid ensemble model" (HEM), which does not require any parameter adjusting. It outperforms other state-of-the-art EL approaches on the majority of performance measures (accuracy, AUC, F1-score). In the absence of prior information on a topic from an internet domain, this study advocates using the suggested HEM model as a generalised learner for binary classification with PCA and IsoMap and for multiclass classification using IsoMap as a feature reduction strategy.

REFERENCES

[1] L. Atzori, A. Iera, and G. Morabito (2010) "The internet of things: A survey," Computer networks, vol. 54, no. 15, pp. 2787–2805, 2010.

[2] M. S. Mahdavinejad (2018) "Machine learning for internet of things data analysis: A survey," Digital Communications and Networks, vol. 4, no. 3, pp. 161–175, 2018.

[3] J. Gubbi, R. Buyya, S. Marusic, and M. Palaniswami (2013) "Internet of things (iot): A vision, architectural elements, and future directions," Future generation computer systems, vol. 29, no. 7, pp. 1645–1660, 2013.

[4] M. Marjani (2017) "Big iot data analytics: architecture, opportunities, and open research challenges," ieee access, vol. 5, pp. 5247–5261, 2017.

[5] A. Janusz, M. Grzegorowski, (2017) "Predicting seismic events in coal mines based on underground sensor measurements," Engineering Applications of Artificial Intelligence, vol. 64, pp. 83–94, 2017.

[6] Z. Geng, H. Wang, (2019) "Predicting seismic-based risk of lost circulation using machine learning," Journal of Petroleum Science and Engineering, vol. 176, pp. 679–688, 2019.

[7] M. J. Ismail, R. Ibrahim, and I. Ismail (2011) "Adaptive neural network prediction model for energy consumption," in 2011 3rd International Conference on Computer Research and Development, vol. 4. IEEE, 2011, pp. 109–113.

[8] Y. K. Penya, C. E. Borges, D. Agote, and I. Fernández (2011) "Short-term load forecasting in air-conditioned non-residential buildings," in 2011 IEEE International Symposium on Industrial Electronics. IEEE, 2011, pp. 1359–1364.

[9] Kubler S., Yoo M. J., Cassagnes C., Framling K., Kiritsis D. and Skilton M. (2015). Opportunity to Leverage Information-as-an-Asset in the IoT-The Road Ahead. in Proceedings - 2015 International Conference on Future Internet of Things and Cloud, FiCloud 2015 and 2015 International Conference on Open and Big Data, OBD 2015, 64–71. doi:10.1109/FiCloud.2015.63.

[10] Andre N., Cho H. W., Baek S. H., Jeong M. K. and Young T. M. (2008). Conformational analysis of a dinucleotide photodimer with the aid of the genetic algorithm. Hoboken, NJ, USA: John Wiley & Sons, Inc. doi:10.1002/0471725382.

[11] Markou M. and Singh S. (2003). Novelty detection: A review - Part 1: Statistical approaches. Signal Processing 83, 2481–2497. doi:10.1016/j.sigpro.2003.07.018.

[12] Chandola V., Banerjee A, and Kumar V. (2009). Anomaly detection: A survey. ACM Comput. Surv. 41, 1–58. doi:10.1145/1541880.1541882.

[13] Worden K., Manson G. and Fieller N. R. J. (2000) Damage detection using outlier analysis. J. Sound Vib. 229, 647–667. doi:10.1006/jsvi.1999.2514.

[14] Filev D. P., Chinnam R. B., Tseng F. and Baruah P. (2010) An industrial strength novelty detection framework for autonomous equipment monitoring and diagnostics. IEEE Trans. Ind. Informatics 6, 767–779. oi:10.1109/TII.2010.2060732.

[15] Kingma D. P., Rezende D. J., Mohamed S. and Welling M. (2014). "Semi-supervised learning with deep generative models," in Advances in Neural Information Processing Systems, eds. Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence and K. Q. Weinberger (Curran Associates, Inc.), 3581–3589. Available at: http://papers.nips.cc/paper/5352- semi-supervised-learning-with-deep-generative-models.pdf

[16] Zeadally S., Adi E., Baig Z., and Khan I. (2020). Harnessing Artificial Intelligence Capabilities to Improve Cybersecurity. IEEE, 8, 1–1, doi.org/10.1109/access.2020.2968045

[17] Kong L. J. (2013). An improved information-security risk assessment algorithm for a hybrid model. International Journal of Advancements In Computing Technology, 5(2).

[18] Liu X., Lu R., Ma J., Chen L. & Qin B. (2016). Privacy-Preserving Patient-Centric Clinical Decision Support System on Naïve Bayesian Classification. IEEE Journal of Biomedical and Health Informatics, 20(2), 655–668. doi.org/10.1109/jbhi.2015.2407157

[19] Shah S. K., Tariq Z., Lee J., & Lee Y. (2021). Event-Driven Deep Learning for Edge Intelligence (EDL-EI). Sensors, 21(18), 6023. doi.org/10.3390/s21186023

[20] Gupta C., Johri I., Srinivasan K., Hu Y.-C., Qaisar S. M., & Huang K.-Y. (2022). A Systematic Review on Machine Learning and Deep Learning Models for Electronic Information Security in Mobile Networks. Sensors, 22(5), 2017. doi.org/10.3390/s22052017

[21] Berman D. S., Buczak, A. L., Chavis J. S. and Corbett C. L. (2019). "Survey of Deep Learning Methods for Cyber Security", Information 2019, 10, 122; doi:10.3390/info10040122

[22] Sarker, I. H., Kayes, A. S. M., Badsha, S., Alqahtani, H., Watters, P., & Ng, A. (2020). Cybersecurity data science: an overview from machine learning perspective. Journal of Big Data. doi.org/10.1186/s40537-020-00318-5

[23] Bringas P. B. and Santos, I., (2010). Bayesian Networks for Network Intrusion Detection, Bayesian Network, Ahmed Rebai, ISBN: 978-953-307-124-4, www.intechopen.com/books/bayesian-network/bayesiannetworks-for-network-intrusion-detection

[24] Bloice, M. & Holzinger, A., 2018. A Tutorial on Machine Learning and Data Science Tools with Python. Graz, Austria: s.n

[25] Wilson, B. M. R., Khazaei, B., & Hirsch, L. (2015, November). Enablers and barriers of cloud adoption among Small and Medium Enterprises in Tamil Nadu. In: 2015 IEEE International Conference on Cloud Computingin Emerging Markets (CCEM) (pp. 140-145). IEEE.

[26] Hashem, I. A. T., Yaqoob, I., Anuar, N. B., Mokhtar, S., Gani, A., & Ullah Khan, S. (2015). The rise of "big data" on cloud computing: Review and open research issues. In Information Systems. https://doi.org/10.1016/j.is.2014.07.006.

[27] Siti Nurul Mahfuzah, M., Sazilah, S., & Norasiken, B. (2017). An Analysis of Gamification Elements in Online Learning to Enhance Learning Engagement. 6th International Conference on Computing & Informatics

[28] R. Baeza-Yates, B. Ribeiro-Neto et al.(1999) Modern information retrieval. ACM press New York, 1999, vol. 463.

[29] T. A. Lasko, J. G. Bhagwat, K. H. Zou, and L. Ohno-Machado (2005) "The use of receiver operating characteristic curves in biomedical informatics," Journal of biomedical informatics, vol. 38, no. 5, pp. 404–415, 2005.

[30] A. Ben-David, (2007) "A lot of randomness is hiding in accuracy," Engineering Applications of Artificial Intelligence, vol. 20, no. 7, pp. 875–885, 2007.

[31] J. Huang and C. X. Ling (2005) "Using auc and accuracy in evaluating learning algorithms," IEEE Transactions on knowledge and Data Engineering, vol. 17, no. 3, pp. 299–310, 2005.

[32] H. Yin and N. K. Jha (2017) "A health decision support system for disease diagnosis based on wearable medical sensors and machine learning ensembles," IEEE Transactions on Multi-Scale Computing Systems, vol. 3, no. 4, pp. 228–241, 2017.

[33] A. O. Akmandor and N. K. Jha (2017) "Keep the stress away with soda: Stress detection and alleviation system," IEEE Transactions on Multi-Scale Computing Systems, vol. 3, no. 4, pp. 269–282, 2017.

[34] J. R. Ng, J. S. Wong, V. T. Goh, W. J. Yap, T. T. V. Yap, and H. Ng, (2019) "Identification of road surface conditions using iot sensors and machine learning," in Computational Science and Technology. Springer, 2019, pp. 259–268.

[35] N. Dogru and A. Subasi, (2008) "Traffic accident detection using random forest classifier," in 2018 15th learning and technology conference (L&T). IEEE, 2018, pp. 40–45.

[36] H. Nguyen, C. Cai, and F. Chen (2017) "Automatic classification of traffic incident's severity using machine learning approaches," IET Intelligent Transport Systems, vol. 11, no. 10, pp. 615–623, 2017.

[37] A. Asuncion and D. Newman (2007) "Uci machine learning repository," 2007.

[38] R. Kohavi et al. (1995) "A study of cross-validation and bootstrap for accuracy estimation and model selection," in Ijcai, vol. 14, no. 2. Montreal, Canada, 1995, pp. 1137–1145.

[39] H. Habibzadeh, A. Boggio-Dandry, Z. Qin, T. Soyata, B. Kantarci, and H. T. Mouftah (2018) "Soft sensing in smart cities: Handling 3vs using recommender systems, machine intelligence, and data analytics," IEEE Communications Magazine, vol. 56, no. 2, pp. 78–86, 2018.

[40] D. J. Sheskin (2020) Handbook of parametric and nonparametric statistical procedures. Chapman and Hall/CRC, 2020.

[41] J. Demšar (2006) "Statistical comparisons of classifiers over multiple data sets," The Journal of Machine Learning Research, vol. 7, pp. 1–30, 2006.

[42] M. Friedman (1937) "The use of ranks to avoid the assumption of normality implicit in the analysis of variance," Journal of the american statistical association, vol. 32, no. 200, pp. 675–701, 1937.

[43] Friedman (1940) "A comparison of alternative tests of significance for the problem of m rankings," The Annals of Mathematical Statistics, vol. 11, no. 1, pp. 86–92, 1940.

[44] R. L. Iman and J. M. Davenport (1980) "Approximations of the critical region of the fbietkan statistic," Communications in Statistics-Theory and Methods, vol. 9, no. 6, pp. 571–595, 1980.

[45] O. J. Dunn, (1961) "Multiple comparisons among means," Journal of the American statistical association, vol. 56, no. 293, pp. 52–64, 1961.

[46] M. Jethanandani, A. Sharma, T. Perumal, and J.-R. Chang (2020) "Multi-label classification based ensemble learning for human activity recognition in smart home," Internet of Things, vol. 12, p. 100324, 2020.

[47] D. García-Gil, J. Luengo, S. García, and F. Herrera (2019) "Enabling smart data: noise filtering in big data classification," Information Sciences, vol. 479, pp. 135–152, 2019.

[48] M. M. Ghiasi and S. Zendehboudi (2021) "Application of decision tree-based ensemble learning in the classification of breast cancer," Computers in Biology and Medicine, vol. 128, p. 104089, 2021.

Table 2: Average Ranks (using Accuracy) of Algorithms on Binary Datasets Using HOV

Sr. No.	Data Sets	LR	KNN	SVM	GB	DT	RF	MLP	NB	AB	QDA	GP
1	Extra Sensory-B	0.797 (6)	0.776 (8)	0.76 (10)	0.87 (1)	0.81 (4)	0.857 (3)	0.775 (9)	0.796 (7)	0.861 (2)	0.806 (5)	0.702 (11)
2	EEG Signal	0.86 (9)	0.87 (6)	0.855 (10)	0.95 (1)	0.912 (3)	0.949 (2)	0.884 (4)	0.863 (8)	0.875 (5)	0.865 (7)	0.836 (11)

3	Churn Modelling	0.788 (8)	0.755 (11)	0.795 (5)	0.858 (3)	0.787 (9)	0.861 (1)	0.795 (5)	0.78 (10)	0.859 (2)	0.832 (4)	0.793 (7)
4	Card Manufacture	0.825 (5)	0.695 (9)	0.57 (10)	0.841 (3)	0.829 (4)	0.856 (1)	0.754 (8)	0.791 (6)	0.842 (2)	0.786 (7)	0.565 (11)
5	Dota Games	0.521 (7)	0.513 (8)	0.525 (4)	0.548 (1)	0.509 (9)	0.533 (3)	0.509 (9)	0.524 (5)	0.539 (2)	0.522 (6)	0.502 (11)
6	Electric Grid	0.891 (9)	0.789 (11)	0.901 (8)	1.0 (1)	1.0 (1)	1.0 (1)	0.921 (7)	0.978 (5)	1.0 (1)	0.967 (6)	0.812 (10)
7	Indian Liver Patient	0.704 (2)	0.672 (6)	0.711 (1)	0.696 (5)	0.639 (9)	0.704 (2)	0.642 (8)	0.557 (10)	0.699 (4)	0.553 (11)	0.652 (7)
8	Irish transition	0.721 (11)	0.784 (10)	0.812 (9)	1.0 (1)	1.0 (1)	0.997 (4)	0.937 (5)	0.85 (7)	1.0 (1)	0.855 (6)	0.844 (8)
9	Watch Sensors	0.803 (2)	0.629 (9)	0.664 (8)	0.744 (6)	0.731 (7)	0.8 (3)	0.537 (10)	0.794 (4)	0.811 (1)	0.792 (5)	0.399 (11)
10	OC 01-Driving	0.826 (6)	0.832 (5)	0.826 (6)	0.924 (2)	0.88 (3)	0.926 (1)	0.823 (8)	0.717 (11)	0.843 (4)	0.765 (10)	0.77 (9)
11	OC 01-Road	0.979 (11)	0.987 (8)	0.982 (10)	0.999 (1)	0.998 (2)	0.998 (2)	0.986 (9)	0.995 (6)	0.998 (2)	0.998 (2)	0.991 (7)
12	PC3	0.89 (2)	0.876 (6)	0.892 (1)	0.869 (7)	0.836 (8)	0.886 (3)	0.767 (9)	0.251 (11)	0.877 (5)	0.437 (10)	0.881 (4)
13	PC4	0.899 (1)	0.858 (6)	0.872 (5)	0.899 (1)	0.857 (7)	0.895 (3)	0.737 (10)	0.851 (8)	0.89 (4)	0.545 (11)	0.846 (9)
14	Diabetes	0.774 (2)	0.728 (7)	0.641 (11)	0.742 (6)	0.706 (8)	0.776 (1)	0.705 (9)	0.76 (3)	0.744 (5)	0.753 (4)	0.681 (10)
15	Pulsar Stars	0.95 (2)	0.935 (7)	0.666 (11)	0.946 (3)	0.922 (8)	0.952 (1)	0.94 (4)	0.922 (8)	0.939 (5)	0.936 (6)	0.908 (10)
16	Seismic Bump	0.924 (6)	0.928 (4)	0.932 (1)	0.919 (7)	0.881 (10)	0.929 (3)	0.897 (9)	0.902 (8)	0.928 (4)	0.349 (11)	0.931 (2)
17	Sonar	0.764 (6)	0.757 (7)	0.55 (11)	0.818 (2)	0.722 (8)	0.812 (3)	0.839 (1)	0.675 (9)	0.793 (4)	0.603 (10)	0.788 (5)
18	Hand recog.	0.598 (8)	0.591 (9)	0.664 (3)	0.68 (1)	0.447 (11)	0.522 (10)	0.667 (2)	0.6 (7)	0.632 (6)	0.636 (5)	0.642 (4)
19	Power System B	0.777 (4)	0.689 (10)	0.663 (11)	0.777 (4)	0.751 (6)	0.799 (2)	0.718 (8)	0.78 (3)	0.821 (1)	0.747 (7)	0.692 (9)
20	Machine Sensor	0.907 (1)	0.784 (9)	0.907 (1)	0.814 (7)	0.803 (8)	0.841 (6)	0.892 (4)	0.682 (10)	0.872 (5)	0.521 (11)	0.907 (1)
	Average Rank	**5.45**	**7.8**	**6.8**	**3.2**	**6.5**	**2.8**	**7**	**7.35**	**3.35**	**7.3**	**7.85**

Fig. 1. *OC: Opel Corsa, DC: Data Cortex (Survey)*

Table 3: Annova Testing

SUMMARY for HOV

Groups	*Count*	*Sum*	*Average*	*Variance*
Logistic Regression	2	10.3	5.15	0.18
Nearest Neighbours (KNN)	2	15.65	7.825	0.00125
Linear SVM	2	14.72	7.36	0.6272
Gradient Boosting	2	6.74	3.37	0.0578
Decision Tree	2	11.88	5.94	0.6272
Random Forest	2	5.34	2.67	0.0338
MLP (ANN)	2	12.23	6.115	1.56645
Naive Bayes	2	13.27	6.635	1.02245
AdaBoost	2	9.35	4.675	3.51125
QDA	2	12.84	6.42	1.5488
Gaussian Process	2	16.85	8.425	0.66125

ANOVA for HOV

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	63.91841	10	6.391841	7.147203	**0.00157**	2.853625
Within Groups	9.83745	11	0.894314			
Total	73.75586	21				

Table 4: Annova Testing

SUMMARY for 5x2CV

Groups	*Count*	*Sum*	*Average*	*Variance*
Logistic Regression	2	10.32	5.16	0.1152
Nearest Neighbours (KNN)	2	16.1	8.05	0.08
Linear SVM	2	15.13	7.565	1.32845
Gradient Boosting	2	6.82	3.41	0.2592
Decision Tree	2	11.94	5.97	0.3698
Random Forest	2	5.23	2.615	0.29645

MLP (ANN)	2	13.26	6.63	0.2048
Naive Bayes	2	12.86	6.43	1.8818
AdaBoost	2	9.02	4.51	2.4642
QDA	2	12.44	6.22	0.9248
Gaussian Process	2	17.08	8.54	1.4112

ANOVA for 5x2CV

Source of Variation	*SS*	*Df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	69.50563	10	6.950563	8.189483	**0.000863**	2.853625
Within Groups	9.3359	11	0.848718			
Total	78.84153	21				

Table 5: Annova Testing

SUMMARY for 10FCV

Groups	*Count*	*Sum*	*Average*	*Variance*
Logistic Regression	2	9.77	4.885	0.02645
Nearest Neighbours (KNN)	2	15.55	7.775	0.10125
Linear SVM	2	14.39	7.195	0.49005
Gradient Boosting	2	7.53	3.765	0.19845
Decision Tree	2	11.6	5.8	0.845
Random Forest	2	5.33	2.665	0.16245
MLP (ANN)	2	13.42	6.71	1.7672
Naive Bayes	2	12.66	6.33	1.5138
AdaBoost	2	9.75	4.875	2.53125
QDA	2	11.95	5.975	2.53125
Gaussian Process	2	17.13	8.565	1.32845

ANOVA for 10FCV

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	60.48804	10	6.048804	5.788027	**0.003823**	2.853625
Within Groups	11.4956	11	1.045055			
Total	71.98364	21				

Table 6 Annova Testing and SUMMARY for 10FCV

Groups	*Count*	*Sum*	*Average*	*Variance*
Logistic Regression	2	9.77	4.885	0.02645
Nearest Neighbours (KNN)	2	15.55	7.775	0.10125
Linear SVM	2	14.39	7.195	0.49005
Gradient Boosting	2	7.53	3.765	0.19845
Decision Tree	2	11.6	5.8	0.845
Random Forest	2	5.33	2.665	0.16245
MLP (ANN)	2	13.42	6.71	1.7672
Naive Bayes	2	12.66	6.33	1.5138
AdaBoost	2	9.75	4.875	2.53125
QDA	2	11.95	5.975	2.53125
Gaussian Process	2	17.13	8.565	1.32845

ANOVA for 10FCV

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	60.48804	10	6.048804	5.788027	**0.003823**	2.853625
Within Groups	11.4956	11	1.045055			
Total	71.98364	21				

Table 7 Annova Testing

Summary for Accuracy in Ensemble Models

Groups	*Count*	*Sum*	*Average*	*Variance*
Accuracy	24	2096.905	87.3767	29.9261
AUC	24	21.156	0.8815	0.00372
F1-Score	24	20.274	0.84475	0.00552

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	119737.93	2	59868.96	5999.813	**4.19853E-78**	3.1296
Within Groups	688.51	69	9.97			
Total	120426.44	71				