# DEVELOPMENT OF DATABASES AND MACHINE LEARNING BASED PREDICTION METHODS FOR PROBLEMS OF BIOMEDICAL INTEREST IN THE CURRENT POST GEOMETRIC -ERA

**Rashmikar Kumar Raju,**
PhD Scholar, Department of CSE, Sunrise University, India.
Director- Product Security, Safety, and Energy Efficiency-Testing & Compliance at Vincular Testing Labs India Pvt. Ltd, Bangalore
Email id: rashimikar@rediffmail.com

**Suneel Gupta,**
Asst. Professor, Department of CSE, Sunrise University, India.
Email id: kommi.suneel@gmail.com

## ABSTRACT

*This study performs comprehensive research to identify study that applied or proposed ML models on disease prediction using electronic health data. Machine-learning (ML) methods have recently attracted great attention and have made significant progress in graph applications. To date, most ML approaches have been evaluated on social networks, but they have not been comprehensively reviewed in the health informatics domain. Herein, a review of ML methods and their applications in the disease prediction domain based on electronic health data is presented in this study from two levels: node classification and link prediction. Commonly used ML approaches for these two levels are shallow embedding and graph neural networks (GNN). We considered journals and conferences from four digital library databases. Based on the identified articles, we review the present status of and trends in graph ML approaches for disease prediction using electronic health data. Though the disease prediction field using ML techniques is still emerging, GNN-based models have the potential to be an excellent approach for disease prediction, which can be used in medical diagnosis, treatment, and the prognosis of diseases.*

***Keywords:*** *Machine-learning (ML) methods, electronic health data, GNN-based models, prognosis of diseases, ML techniques.*

## INTRODUCTION

The diversity and the volume of the biological data are increasingly demanding the development of data mining methods with high specificity yet great versatility. Development of machine-learning based prediction method for the diagnosis of head and neck cancer using gene expression profiles. The issues of cost-effectiveness and time consumed in experimental characterization motivate the development of databases and in-silica methods to infer useful biological information from such data. In the foreground of these ideas, we have worked on versatile biomedical problems whereby we developed web accessible databases for four different protein classes- protozoan virulent proteins, lipocalins, fungal adhesins and Cyclin-Dependent Kinase Inhibitors (CDKis). For the latter three protein classes, we also developed machine-learning based prediction methods using two techniques, namely Support Vector Machines (SVM) and Artificial Neural Network (ANN). All of these protein classes are well known for their biomedical merit and relevance to biotechnological and therapeutic applications, yet their identification presents a challenging task owing to the diversity in amino acid sequence. To aid the end-user biologist, the best classifiers obtained in each study have been implemented on the web interface. We also applied SVM and ANN to one microarray

dataset from NCBI Gene Expression Omnibus (GEO) for the diagnosis of head and neck cancer and obtained promising accuracies. We have developed machine learning classifiers as a complementary tool to the already existing ones, and validate its efficiency on independent datasets. In this work, we collate all such information and present it in the form of a publicly accessible user-friendly database, besides integrating several useful comparative tools with it.

## LITERATURE REVIEW

**Abdullah Al Foysal (2025)** Mental disorders, including depression, bipolar disorder, and mood disorders, affect millions of individuals worldwide, significantly impacting their quality of life. Early and accurate diagnosis is essential for effective intervention, reducing the burden on healthcare systems, and improving patient outcomes. However, traditional diagnostic methods rely heavily on subjective assessments, self-reported symptoms, and clinical observations, which may lead to delays and inconsistencies in diagnosis. The integration of artificial intelligence (AI) and machine learning (ML) in mental health care has emerged as a promising solution to enhance predictive accuracy and provide early diagnosis. This study explores the application of ML algorithms to predict mental disorders using behavioral and psychological features. The dataset comprises attributes such as sadness, sleep disorders, mood swings, anxiety levels, and suicidal thoughts.

**Manojit Bhattacharya (2024)** The medicine and healthcare sector has been evolving and advancing very fast. The advancement has been initiated and shaped by the applications of data-driven, robust, and efficient machine learning (ML) to deep learning (DL) technologies. ML in the medical sector is developing quickly, causing rapid progress, reshaping medicine, and improving clinician and patient experiences. ML technologies evolved into data-hungry DL approaches, which are more robust and efficient in dealing with medical data. This article reviews some critical data-driven aspects of machine intelligence in the medical field. Here, we discuss the progress of ML, DL, and the transition requirements from ML to DL. To discuss the advancement in data science, we illustrate prospective studies of medical image data, newly evolved DL interpretation data from EMR or EHR, big data in personalized medicine, and dataset shifts in artificial intelligence (AI).

**Karen Gishyan (2023)** Finding out the desired drug combinations is a challenging task because of the number of different combinations that exist and the adversarial effects that may arise. In this work, we generate drug combinations over multiple stages using distance calculation metrics from supervised learning, clustering, and a statistical similarity calculation metric for deriving the optimal treatment sequences. The combination generation happens for each patient based on the characteristics (features) observed during each stage of treatment. Our approach considers not the drug-to-drug (one-to-one) effect, but rather the effect of group of drugs with another group of drugs. We evaluate the combinations using an FNN model and identify future improvement directions.

**Barnabas Achakpa Ikyo (2021)** The ability of machine learning techniques to make accurate predications is increasing. The aim of this work is to apply machine learning techniques such as Support Vector Machine, Naïve Bayes, Decision Tree,

Logistic Regression, and K-Nearest Neighbour algorithms to predict the shelf life of Okra. Predicting the shelf life of Okra is important because Okra becomes harmful for human consumption if consumed after its shelf life. Okra parameters such as weight loss, firmness, Titrable Acid, Total Soluble Solids, Vitamin C/Ascorbic acid content, and PH were used as inputs into these machine learning techniques. Support Vector Machine, Naïve Bayes and Decision Tree each accurately predicted the shelf life of Okra with accuracies of 100%. However, the Logistic Regression and K-Nearest Neighbour achieved 88.89% and 88.33% accuracies, respectively.

**Yufa Wang (2018)** The aim of bankruptcy prediction is to help the enterprise stakeholders to get the comprehensive information of the enterprise. Much bankruptcy prediction has relied on statistical models and got low prediction accuracy. However, with the advent of the AI (Artificial Intelligence), machine learning methods have been extensively used in many industries (e.g., medical, archaeological and so on). In this paper we compare the statistical method and machine learning method to predict bankruptcy with utilizing China listed companies. Firstly, we use statistical method to choose the most appropriate indicators. Different indicators may have different characteristics and not all indicators can be analyzed. After the data filtering, the indicators are more persuasive. Secondly, unlike previous research methods, we use the same sample set to conduct our experiment. The final result can prove the effectiveness of the machine learning method.

**Machine Learning Prediction**

Machine learning prediction, or prediction in machine learning, refers to the output of an algorithm that has been trained on a historical dataset. The algorithm then generates probable values for unknown variables in each record of the new data. The purpose of prediction in machine learning is to project a probable data set that relates back to the original data. This helps organizations predict future customer behaviors and market changes. Essentially, prediction is used to fit a shape as closely to the data as possible. With machine learning predictions, organizations use proactive decisions to avoid predicted user churn. To gain the most success with prediction in machine learning, organizations need to have infrastructure in place to support the solutions, and high-quality data to supply the algorithm.

**Examples of Machine Learning Prediction**

Prediction can be used to forecast the future and to predict the probability of an outcome. It can also be used to forecast future requirements or run a what-if analysis. One prediction tool is regression analysis which is used to determine the relationship between two variables (single regression) or more than two variables (multiple regression). Predictive analytics is when data is used to predict future trends or events. With predictive analytics, historical data is used to forecast potential scenarios and use these predictions to drive strategic business aimed decisions. Prediction can also be used to forecast future cash flow, determine staffing needs in the hospitality and entertainment industry, predict user behavior, prevent malfunctioning, and predict potential allergic reactions for patients in the healthcare industry.

**Machine Learning Prediction Important**

Prediction in machine learning allows organizations to make predictions about possible outcomes based on historical data. These assumptions allow the organization to make decisions resulting in tangible business results. Predictive analytics can be used to anticipate when users will churn or leave an organization. With this recognition, organizations have better potential to keep customers happy and satisfied.

**Use of Prediction**

The prediction audience includes people who are in need of answers to future questions in order to make business decisions. Predictive analysis provides confident answers to complicated problems, explores new kinds of problems, and offers real time answers to problems with changing information. Prediction can be used by everyone because of its vast capabilities. It is helpful for businesses and other organizations when making decisions, but can also be used to make movie recommendations. Prediction is also used to detect fraud in previous transactions which is a function often used by banks.

**Prediction vs Traditional Methods**

Machine learning prediction is preferred over traditional methods because it is usually a better predictor. Since machine learning uses algorithms, it can identify patterns and relationships that humans cannot. Larger data sets are also able to be analysed and turned into predictions. Traditional methods use humans for computation, which requires more time, money, and is subject to bias by human emotion or opinion. With users and the market constantly changing, machine learning prediction also provides the benefit of adapting quickly and higher efficiency.

**Prediction vs classification:** Classification is separating data into classes, whereas prediction is about fitting a shape that gets as close to the data as possible.

**Prediction vs inference statistics:** Prediction is the process of a machine learning model predicting potential data points. Inference statistics evaluate the difference between predictor and response variables.

**Databases for Machine Learning**

Databases are a critical element in machine learning today. It helps you train various machine learning and artificial intelligence (AI) models. The excellent benefits that these technologies offer are the primary reason behind their growing use of this technology. In the past few decades, many new datasets have been available. As a result, it might be a challenge to choose the best one for your tasks. However, it also allows businesses to choose from the large number of datasets that can be the perfect fit for the application plan. So, what are the best databases for machine learning that you can find in the market? Should you go for a free AI database or a customized one? And what is the advantage of using customized databases for your ML tasks? We'll discuss all those things in this study.

**Best Databases for Machine Learning and Artificial Intelligence**

Choosing the correct databases for your machine learning and Artificial Intelligence tasks can ensure you get the desired results. We have listed the top ten databases and their core features to make things easy. You can choose any one of them according to your needs.

Redis: Redis is a top-notch open-source, in-memory data structure many people currently use in the market. You can use it

as a database for machine learning and AI projects or tasks.

The best thing about Redis is that it supports various data structures like bitmaps, geospatial indexes, sorted sets, etc. Additionally, you can also find the following features if you choose Redis as a database:

- Transactions Lua scripting LRU eviction
- Different levels of on-disk persistence
- Built-in replication

It also comes with an automatic failover process. You can also use Redis to write complicated code with fewer and easy lines. So, if you are looking for a robust database for your machine-learning tasks, then Redis is an optimal choice.

**RESEARCH METHODOLOGY**

The first step for developing any prediction method is the collection of good quality and quantity of data. For protein sequence data, 'good quality' implies the experimental validation and/or availability of good annotation about the sequences. The performance of the most efficient classifiers is further checked on independent datasets to judge their utility for practical application. Selection of the appropriate database for either sequence or microarray data is of prime importance, as all training examples should have unambiguous annotation and/or experimental evidence for the desired function/property. Moreover, the development of these methods requires the preparation of both positive and negative datasets. The negative set is more often prepared by randomly selecting sequences from common databases. However, translation of training dataset into higher-dimensional space incurs both computational costs and tendency of finding trivial solutions; leading to over-fitting of data. The best parameters as measured by the various performance measures are picked up and then averaged for the fmal assessment of the model. For microarray data, it refers to the quality of the initial data as well as appropriate noise reduction pre-processing of the data. 'Good quantity' implies that the number of samples should be sufficiently high to develop a classifier with good efficiency.

**RESULTS AND DISCUSSIONS**

The three panels were then pooled together to identify 42 probe sets (representing 38 genes and one expressed sequence tag) representing differentially expressed genes (Table 1). Selected probe sets were validated by hierarchical clustering analysis (HCA), multiple probe set concordance, and target-subunit agreement. Moreover, real-time PCR of 8 representatives (randomly selected from 38 genes) performed on both micro-array tested and independently obtained samples correlated well with the microarray data.

**Table 1: List of differentially expressed genes**

| Category | Up-regulated in tumors | | | Down-regulated in tumors | | | Category |
|---|---|---|---|---|---|---|---|
| | probe ID | target | FC* | probe ID | target | FC* | |
| [I] | 35474_s_at | COL1A1 | 2.7 | 39657_at | KRT 4 | -16.9 | [I] |
| Structural | 32305_at | COL1A2 | 3.1 | 36883_at | KRT 13 | -7.6 | see left |
| or | 32306_g_at | COL1A2 | 3.3 | 36890_at | PPL | -4.6 | |
| associated | 39333_at | COL4A1 | 2.6 | 35105_at | SCEL | -5.0 | |
| | 36659_at | COL4A2 | 2.5 | 32200_at | ACPP | -2.4 | |
| | 38420_at | COL5A2 | 2.9 | 617_at | ACPP | -2.6 | |
| | 31719_at | FN1 | 3.5 | 37125_f_at | CYP3A5 | -3.5 | [II] |
| | 38442_at | MFAP2 | 1.9 | 529_at | DUSP5 | -2.3 | Enzyme |
| | | | | 770_at | GPX3 | -2.7 | or |
| [II] | 39945_at | FAPA | 2.9 | 32570_at | HPGD | -3.0 | inhibitor |
| see right | 38428_at | MMP1 | 4.7 | 37093_at | PP11 | -6.3 | |
| | 37310_at | PLAU | 2.8 | 40315_at | SPINK5 | -9.7 | |
| | 39166_s_at | SERPINH2 | 2.0 | 32868_at | TGM3 | -15.9 | |
| | | | | 988_at | CEACAM1 | -4.3 | [III] |
| [III] | 2092_s_at | OPN | 5.8 | 1582_at | CEACAM5 | -3.9 | Tumor- |
| see right | 34342_s_at | OPN | 6.0 | 39698_at | LAGY | -6.6 | associated |
| | | | | 1321_s_at | EMP1 | -3.9 | antigen |
| [IV] | 39710_at | C5ORF13 | 1.9 | 38242_at | BLNK | -2.9 | |
| Other | 33410_at | ITGA6 | 2.8 | 37603_at | IL1RN | -4.9 | |
| | 1837_at | NA | 1.8 | 41644_at | KIAA0790 | -2.3 | [IV] |
| | 615_s_at | PTHLH | 2.5 | 38051_at | MAL | -14.4 | Other |
| | | | | 33483_at | NMU | -3.6 | |
| | | | | 37920_at | PITX1 | -3.2 | |
| | | | | 32139_at | ZNF185 | -3.4 | |

Table 2 and 3 summarize the performances of the various SVM and ANN classifiers respectively. All the SVM classifiers were trained using the linear kernel. All the ANN classifiers were trained using the standard vanilla backpropagation algorithm. We developed three types of SVM and ANN classifiers. First, we used all the 12625 genes from the 22 positive (tumour) and negative (normal) samples. Secondly, we used the 42 differentially expressed genes selected by Kuriakose and colleagues in their work. Thirdly, we performed SVM-RFE to select a small number of genes which gave the maximum accuracy and we obtained 8 genes as a subset of 42 genes. These genes included structural proteins collagen 1A1, microfibrillar-associated protein 2 (MFAP2) and keratin, as well as dual specificity phosphatase 5, matrix metalloproteinase 1, plasminogen activator, glutathione peroxidase and SAM and SH3 domain-containing protein (SASH1).

**Table 2: Performance of various SVM classifiers with the number of genes as shown in the first column and discussed in the text**

| #genes | C | Th | SN (%) | SP(%) | Accuracy(%) | MCC |
|---|---|---|---|---|---|---|
| 12625 | 0.1 | -0.2 | 100 | 95.45 | 97.72 | 0.955 |
| 42 | 0 | 0.2 | 100 | 95.45 | 97.72 | 0.909 |
| 8 | 0 | 0.2 | 100 | 100 | 100 | 1.000 |

#genes- Number of genes, C- Regularization parameter for linear kernel, Th- Threshold, SN- sensitivity, SP - specificity, MCC - Matthews Correlation Coefficient,

It is evident that the performance of ANN classifiers progresses from an accuracy of 77.27% to 97.72% to 100% as the number of genes reduces from 12625 to 42 to 8. On the contrary, SVM classifiers show only a marginal increase from 97.72% accuracy to 100% accuracy for the reduction to 8 genes. The accuracy remains stable at 97.72% for the reduction from 12625 genes to 42 genes.

**Table 3: Performance of various ANN classifiers with the number of genes as shown in the first column and discussed in the text**

| #genes | H,cyc | Th | SN (%) | SP(%) | Accuracy(%) | MCC |
|---|---|---|---|---|---|---|
| 12625 | 4, 400 | 0.3 | 77.27 | 77.27 | 77.27 | 0.545 |
| 42 | 2,400 | 0.5 | 95.45 | 100 | 97.72 | 0.955 |
| 8 | 500 | 0.5 | 100 | 100 | 100 | 1.000 |

#genes- Number of genes, Th- Threshold, SN- Sensitivity, SP- Specificity, MCCMatthews Correlation Coefficient, H-number of hidden neurons, eye- number of training cycles

## CONCLUSIONS

An extensive review of existing research was conducted and with practical application of models it was attempted to generate insight on the workings of the methods. However, finding an implementation to train the models as intended, with different settings, with a common evaluation strategy, practical difficulties were faced. Commonly applicable solutions, with a straight forward way of interpretation like for many tasks in vision are not (yet) in existence. Dealing with large-scale heterogeneous graphs, usable methodology is somewhat limited but solutions for these are rapidly multiplying. Further, more telling metrics should be inspected, visualizations of the training on the embeddings could be applied and the predictions using the embeddings could be analyzed regarding their realistic usefulness. A suite of downstream tasks could be created as benchmark, to assess e.g. if the embedding models using a specific similarity definition for embedding are actually suited to capture useful biological knowledge or whether for certain applications rules should be incorporated. Machine learning means that representations can be automatically fit to the data but a perfect model that can be used for most cases is non-existent. Using RFE, we obtained the molecular signature of 8 genes to discriminate perfectly between tumour and normal samples. The number of genes may further reduce with a larger number of samples. The work on RFE represents only preliminary experiments of feature selection and we do not conclude that these are the final markers for the disease.

## REFERENCES

1. *Airin Afroj Aishi (2024), "Machine learning and deep learning-based approach in smart healthcare: Recent advances, applications, challenges and opportunities", AIMS Public Health, issn: 2327-8994, vol. 11(1), pages. 58-109.doi: 10.3934/publichealth.2024004*
2. *Manojit Bhattacharya (2024), "From machine learning to deep learning: Advances of the recent data-driven paradigm shift in medicine and healthcare", Current Research in Biotechnology, issn: 2590-2628, vol. 7, https://doi.org/10.1016/j.crbiot.2023.100164*
3. *Barnabas Achakpa Ikyo (2021), "Application of Machine Learning Techniques for Okra Shelf-Life Prediction", Journal of Data Analysis and Information Processing, issn: 2327-7203, Vol.9, No.3, pages.136-150.*
4. *Yufa Wang (2018), "Machine Learning Methods of Bankruptcy Prediction Using Accounting Ratios", Open Journal of Business and Management, issn: 2329-3292, vol.6, pages.1-20.*
5. *Elchin Asgarov (2024), "A Comprehensive Analysis of Machine Learning Techniques for Heart Disease Prediction", Open Access Library Journal, issn: 2333-9721, vol.11, pages.1-17.*
6. *Karen Gishyan (2023), "Drug-Treatment Generation Combinatorial Algorithm Based on Machine Learning and Statistical Methodologies", Open Journal of Applied Sciences, issn: 2165-3925, vol.13, pages.548-561.*
7. *Abdullah Al Foysal (2025), "AI-Driven Mental Disorder Prediction: A Machine Learning Approach for Early Detection", Open Access Library Journal, issn: 2333-9721, vol.12, pages.1-1.*
8. *Han Zhou (2023), "Research on Dynamic Mathematical Resource Screening Methods Based on Machine Learning", Journal of Applied Mathematics and*

*Physics, issn: 2327-4379, vol.11, pagees.3610-3624.*

9. *Xiao-Jun Zeng (2013), "Evaluation and Comparison of Different Machine Learning Methods to Predict Outcome of Tuberculosis Treatment Course," Journal of Intelligent Learning Systems and Applications, issn: 2150-8410, Vol. 5, No. 3, pp. 184-193.*
10. *Mohammad Zubair Khan (2019), "Software Defect Prediction Using Supervised Machine Learning and Ensemble Techniques: A Comparative Study. Journal of Software Engineering and Applications, issn: 1945-3124, vol.12, pages.85-100.*