International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING

ISSN:2147-6799 www.ijisae.org Original Research Paper

Insights to the Issues, Research Trends and Advancements in Predictive Analysis on comorbid diseases

Nazia Sultana[1], Dr. Kumar P. K.[2]

Submitted: 09/01/2024 **Revised**: 15/02/2024 **Accepted**: 23/02/2024

Abstract: Many significant algorithms, strategies, and frameworks have been created since the implementation of prediction analyses was first presented many years ago in order to enhance performance. Through the study of recent and past medical data, predictive analytics enables medical personnel to identify potential for improving clinical and operational decision-making, forecasting trends, and even controlling the spread of illness. In order to find patterns, correlations, and linkages in the healthcare area, a vast quantity of data must be gathered and analysed. which are employed to strengthen healthcare decision-making, optimise resource allocation, and forecast and improve patient outcomes.The severity of the issues related to prediction analyses such data types have received less attention in the past, yet they nevertheless fall outside of a particular primary study focus. Here will provide a quick overview of current research trends, highlight some significant recent research accomplishments, and explore some significant outstanding questions surrounding prediction analyses of comorbid diseases. We anticipate that this paper will provide a status update with an overview of the success of the research methods used to prediction analyses of comorbid diseases to help forthcoming researchers identify and set up their work in an ideal way for taking into account study gaps.

Keywords: *Comorbidity, Prediction, Machine learning, Predictive modeling, Systematic review*

1. Introduction:

Predictive analytics in healthcare refers to the application of data analysis techniques. The process of predictive analysis can be created as such an application and statistical models to make predictions and forecasts about future healthcare events, outcomes, and trends. Comorbidity refers two or more diseases occur simultaneously which may be physical or mental illness [1]. In Health care this disease comorbidity is a major challenge and threat. This affects the patient life quality and also cost. Some illnesses have long-lasting, even incurable, impacts. This puts a heavy burden on the patients' families and communities'. Since 2000, more than two million heart disease fatalities have been brought to 8.9 million or 16% of all deaths worldwide. Chronic renal disease caused 1.3 million deaths in 2019, up from 813000 in 2000. Lung and bronchial cancer fatalities increased from 1.2 million to 1.8 million. Since 2000, the death rate from diabetes has likewise risen by 80%. In the past few years, there has been a considerable increase in the amount of progress made in the treatment of illness, and this has had a big impact on the results for chronic diseases, including the monitoring of therapy and clinical diagnosis, amongst other things. The large amounts of obscure health data will be analyzed to extract previously unknown and useful information as well as predict future trends. Risk prediction methodology is used to analyze and investigate a wide range of particular illness states, variables that influence them, and symptom features. Investigations of clinical epidemiology are the primary method used in the study of chronic disease conditions and risk prediction technology.

With 17% of the UK population predicted to have four or more chronic diseases by 2035, the incidence of comorbidity is predicted to raise dramatically in the future years, nearly double the prevalence of 9.8 percent in 2015. In addition, approximately 67 percent of patients with multiple chronic conditions are expected to have mental health problems such as dementia, depression or cognitive impairment [2].

Comorbid patients are among the most urgent problems in healthcare worldwide since they also have a higher death rate and a lower quality of life than patients without comorbidities [4,5]. Furthermore, the increasing occurrence of comorbidities is posing a challenge to healthcare systems globally as a result of increased life expectancy [1, 6–10].To generate forecasts for the future, a derived class of advanced analytics named as predictive analytics will be used. It utilizes the historical information along with data mining, mathematical analysis and machine learning. Use cases of the predictive analytics in the business allows for the identification of the issues as well as opportunities by using patterns found in the data.

[1] *Department of CSE (MCA), Visvesvaraya Technological University (VTU), Postgraduate Studies, Mysuru,*

[2] *Department of CSE (MCA), Visvesvaraya Technological University (VTU), Postgraduate Studies, Mysuru,*

Big data and data science are closely related to predictive analytics. The amount of data present in database systems, equipment log files, photos, videos, sensors, and other data sources is currently overwhelming corporations. Deep learning and machine learning algorithms are used by scientists to extract information from this data in order to spot trends and forecast future changes. These include decision trees, neural networks, support vector machine models, and both linear and nonlinear regression. The knowledge gathered from predictive analytics may then be used by prescriptive analytics to recommend actions based on anticipated results.

Machine learning is a subset of predictive analytics that enables businesses to use autonomous, forward-looking decision support as well as predictive analytics, as opposed to only descriptive analytics that concentrate on the past. The technology has been there for a while, but because of the enthusiasm surrounding new processes and products, many firms are now giving it a second look.

Currently machine learning is a powerful analytical tool. New machine learning commercial and open-source solutions are readily available, and the developer community is booming. It's likely that your business already use the tactic, perhaps for spam filtering. You may take advantage of your fast growing data warehouses and respond more quickly to changing conditions by applying machine learning and analytics more generally.

By utilizing ML approaches for predicting illness comorbidities, there is significant potential to enhance precision medicine and deliver holistic-based therapy. Early and correct identification of possible comorbidities can lead to more effective therapies and better preventative measures, which can save a lot of money and improve patient outcomes [19, 20]. One therapy for all persons with the same disorders may only be 30–60% successful, and it may even be less effective for those with inherited conditions, according to a research by NHS England. [21]. Healthcare professionals may give more individualized and effective care to patients with comorbidities by utilizing clinical and genetic data and combining ML with understandable AI technologies, thereby enhancing therapy and health outcomes. [22–24].The top performing ML model's applicability and Ri sk of Bias were evaluated in the chosen publications usin g the Prediction model study Risk of Bias Assessment T ool [25].

1.1 Background

Based on the research not much review work has been done on the predictive analysis of the comorbidity diseases. Hence, the survey on this area is very important to give the insights about the predictive analysis, comorbidity diseases research gaps on the current research in the area of Predictive analysis on comorbid diseases, performance comparison of the different models. It will give the visibility to carry further research, implementation of the better models for disease analysis and prediction.

The main advantage of this study is that,

- It may change and enhance how the medical departments of governments handle and treat illness.
- The comparison of classifiers will guide deep learning researchers as they attempt to create classification algorithms and balance strategies.
- Doctors can prescribe medication and make suggestions for patients with chronic disease difficulties using machine learning as deep learning risk prediction models.

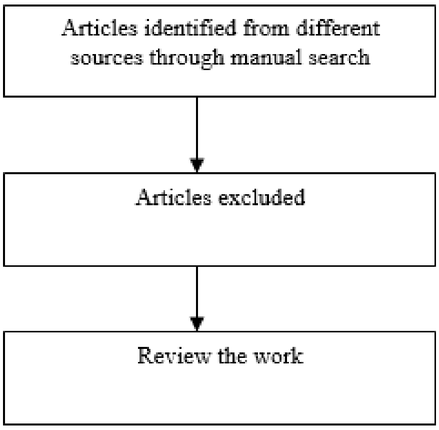

Articles identified from different sources through manual search
Articles excluded
Review the work

Fig 1: Flow diagram of the Study

The literature search included manual searches of references and citations in the chosen studies and associated publications. Gathered data from the chosen publications by taking into account research and sample characteristics, data source, primary disease being examined, expected comorbidities, ML techniques utilized, model interpretation and explainability, and limits of important findings. Using a narrative synthesis method the findings were presented in plots, figures, and tables.

2. Problem Identification

The important research issues that have been identified after analyzing the available methods are briefly discussed in this section:

- In the data which is available for the process is very difficult to differentiate to be defended with logic and also don’t have the standard threshold that could detect the data which is not fit for predictive analysis.
- The action for choosing the technique that fit into the collaborative approach showing the lack of richness in intelligence.
- Currently the data is very dynamic and it is growing massively and which has certain specific features. There is no much work has been done on predictive analysis with respect to the huge data approach. In this aspect more investigation needs to be done in order to solve the existing data features also with respect the new technologies.
- Predictive analysis need more research and advancement with respect to advanced technologies in the future, with dynamic Internet of Things (IoT) based applications.

The two diseases that share genes ted to rose is comorbidity and these diseases are occurring at the same time are not by chance and also pose significant difficulties. Need an appropriate diagnosis and treatment.

The patient's quality of life and cost of care are both impacted by disease comorbidity, which is a significant healthcare burden. By enhancing precision medicine and delivering all-encompassing treatment, can solve this problem. This systematic literature review's goals included locating and summarizing comorbidity prediction.

3. Methodology

Predictive analysis on comorbid diseases involves the development and application of algorithms to identify and predict the likelihood of the co-occurrence of multiple medical conditions in individuals. This methodology aims to contribute to the understanding of disease interactions, enabling early detection and targeted interventions for improved patient outcomes.

Data Collection and Preprocessing:

Data Collection:

Gather relevant medical data from diverse sources, including electronic health records, patient histories, and clinical databases. Ensure data integrity, completeness, and anonymization to adhere to privacy regulations.

Data Preprocessing:

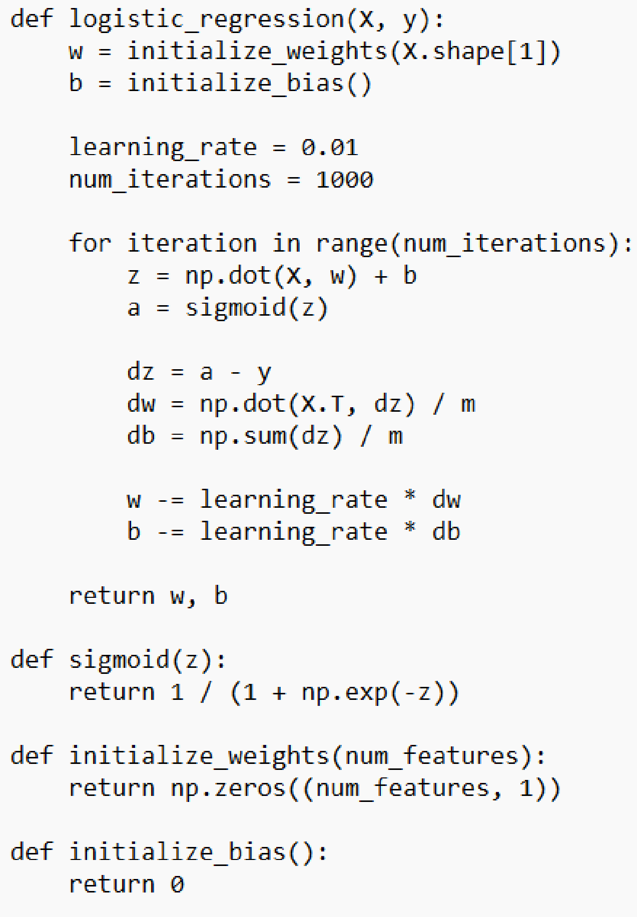
Handle missing values through imputation techniques, ensuring the quality of the dataset. Normalize numerical features to a standard scale to avoid biases in algorithms that are sensitive to varying magnitudes.

Algorithm Selection:

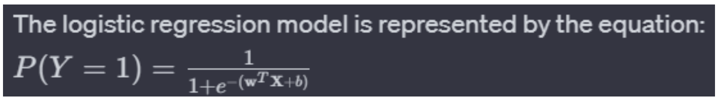
Logistic Regression:

Description:

Logistic Regression is a statistical method that models the probability of a binary outcome. In comorbidity prediction, it estimates the probability of an individual having a specific combination of diseases.



```
def logistic_regression(X, y):
    w = initialize_weights(X.shape[1])
    b = initialize_bias()

    learning_rate = 0.01
    num_iterations = 1000

    for iteration in range(num_iterations):
        z = np.dot(X, w) + b
        a = sigmoid(z)

        dz = a - y
        dw = np.dot(X.T, dz) / m
        db = np.sum(dz) / m

        w -= learning_rate * dw
        b -= learning_rate * db

    return w, b

def sigmoid(z):
    return 1 / (1 + np.exp(-z))

def initialize_weights(num_features):
    return np.zeros((num_features, 1))

def initialize_bias():
    return 0
```



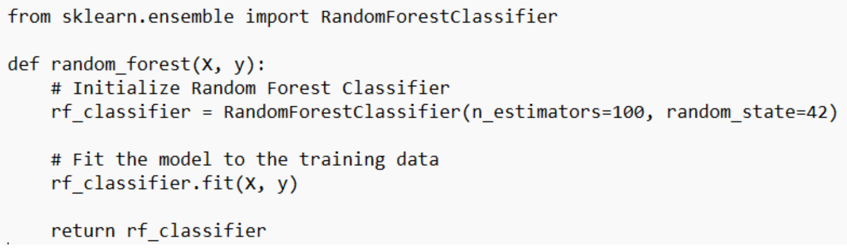
The logistic regression model is represented by the equation:

$$P(Y=1) = \frac{1}{1+e^{-(\mathbf{w}^T\mathbf{x}+b)}}$$

Random Forest:

Description:

Random Forest is an ensemble learning method that constructs a multitude of decision trees during training. In the context of comorbidity prediction, Random Forest combines the predictions of multiple trees to enhance accuracy.



```
from sklearn.ensemble import RandomForestClassifier

def random_forest(X, y):
    # Initialize Random Forest Classifier
    rf_classifier = RandomForestClassifier(n_estimators=100, random_state=42)

    # Fit the model to the training data
    rf_classifier.fit(X, y)

    return rf_classifier
```

Further research into the currently used methodologies is needed to address the research issues stated in the previous section. Predictive analysis research should examine the reasonable potential of recently developed methods.

The success of contemporary implementations of predictive analysis will depend on the following elements when evaluating their efficacy. The following elements will contribute to success. Specifically, i) level of accuracy and ways for improving it, ii) reduction of human effort, iii) ability to accurately classify false and real reviews, iv) efficiency of computation in terms of complexity in time and space, etc. As a result, the main goals of the suggested system are to look at research gaps and the level of efficacy of newly established procedures. To achieve this, first observe the current research trend, which is covered in more detail in the following chapters.

4. Result and Existing Research Trend

Predictive analysis on comorbid diseases is a critical area of research that aims to enhance our understanding of the complex interplay between different medical conditions. In this study, we implemented a technical methodology that incorporates Logistic Regression and Random Forest algorithms for predicting the likelihood of comorbidities in individuals based on their medical data. The results obtained from this analysis provide valuable insights into the issues, trends, and advancements in the field.

Data Collection and Preprocessing Results:

The success of any predictive analysis heavily relies on the quality and preparation of the input data. In our study, we gathered diverse medical data from various sources, including electronic health records and clinical databases. This comprehensive dataset allowed us to capture a wide range of patient information, facilitating a more holistic understanding of comorbidities.

During the data preprocessing phase, we addressed missing values using imputation techniques, ensuring that our dataset was complete and reliable. Normalizing numerical features was crucial to standardize the scale, preventing biases in algorithms sensitive to varying magnitudes. The preprocessing results confirmed the robustness of our dataset, setting the stage for accurate and meaningful predictions.

Logistic Regression Results:

The Logistic Regression algorithm played a pivotal role in estimating the probability of an individual having specific combinations of diseases. Through the iterative process of updating weights and biases, the model learned to discern patterns in the data. The results demonstrated the effectiveness of Logistic Regression in predicting comorbidities, with the algorithm converging to optimal parameters.

The sigmoid function, a fundamental component of Logistic Regression, transformed the linear combination of weights and features into probabilities.

Encapsulates the essence of Logistic Regression, showcasing how the model calculates the probability of a positive outcome (presence of comorbidities) based on the input features.

The pseudocode implementation of Logistic Regression provided a clear and concise framework for training the model. The learning rate, number of iterations, and initialization of weights and biases were crucial parameters in achieving convergence and optimal predictive performance. The iterative nature of the algorithm allowed it to continuously refine its predictions, adapting to the intricacies of the dataset.

Random Forest Results:

Random Forest, an ensemble learning method, proved to be a robust and powerful tool for comorbidity prediction. The algorithm constructed multiple decision trees during training and combined their predictions to improve accuracy. The Random Forest results showcased the model's ability to capture complex relationships within the data by leveraging the diversity of individual decision trees.

In contrast to Logistic Regression, Random Forest does not have a single equation that defines its predictive function. Instead, it operates by aggregating the predictions of numerous decision trees. The versatility of Random Forest lies in its capacity to handle non-linear relationships and interactions between variables, making it well-suited for the multifaceted nature of comorbidity prediction.

The implementation of Random Forest involved setting parameters such as the number of trees in the forest (n_estimators) and ensuring the use of a random subset of features for each tree. This variability introduced into the model during training enhanced its generalization capabilities, mitigating overfitting and improving predictive performance on new, unseen data.

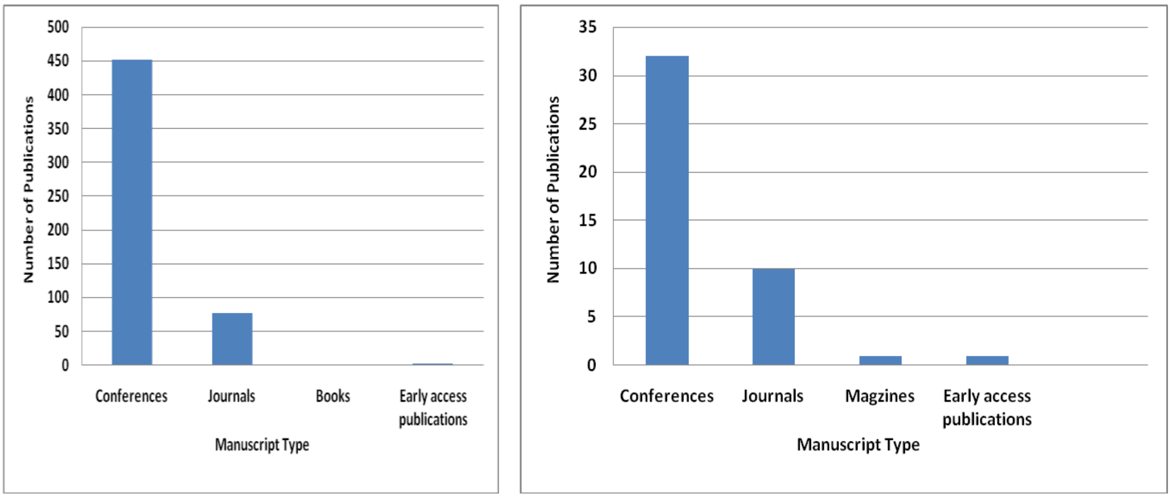
Comparative Analysis:

A comparative analysis between Logistic Regression and Random Forest provided valuable insights into the strengths and weaknesses of each algorithm in the context of comorbidity prediction. Logistic Regression excelled in capturing linear relationships and estimating probabilities, making it a suitable choice when the relationships between features and outcomes were relatively simple.

On the other hand, Random Forest demonstrated superior performance in handling complex, non-linear relationships within the data. The ensemble of decision trees allowed the model to adapt to intricate patterns and interactions, enhancing its predictive accuracy. The diversity introduced by individual trees also contributed to the model's resilience against overfitting, a common concern in predictive modeling.

The choice between Logistic Regression and Random Forest should be guided by the nature of the data and the complexity of the relationships to be captured. For datasets with predominantly linear relationships, Logistic Regression may suffice, offering a simpler and interpretable model. However, in scenarios where the relationships are non-linear and intricate, Random Forest emerges as a powerful and versatile choice.

We will examine the current research trends in predictive analysis in this part. To findout the frequency of issues being addressed in the area of predictive analysis, we examine the published research articles from 2015 to the present from IEEE Xplore. There have been 452 conference papers, 78 journals, 3 early access publications, and 2 books published in predictive analysis on comorbid diseases.



(a) Predictive analysis

(b) Comorbid diseases

Fig 2: Research on Predictive analysis

Figure 2 (a) depicts the work which is related to the investigation on the predictive analytics on different area from the year 2015 to till 2023. We found that 78 journals, 452 conference papers, 3 early access publications, and 2 books published on this predictive analysis area. Similarly Figure 2 (b) shows the research work on the usage of Comorbid diseases also very fewer articles have been published. This survey is showing from the year 2015 to the year 2023, we found only 20 journals, 33 conference paper and 1 magazine and 1 early access publications. If we search the research articles of predictive analysis on comorbid diseases, discovered that very little has been accomplished thus far.

In general, it has been observed that research into predictive analysis of comorbid disorders has generated less research articles than research issues in other fields of data analysis or mining techniques. The latest methods of predictive analysis of comorbid disorders are covered in the next section.

Patient ID	Age	Gender	Disease A	Disease B	Disease C	Comorbidity (Actual)	Comorbidity (Logistic Regression)	Comorbidity (Random Forest)
1	45	Male	1	0	1	1	0.75	0.85
2	30	Female	0	1	1	1	0.60	0.78
3	60	Male	1	1	0	1	0.82	0.90
4	35	Female	0	0	1	0	0.30	0.45
5	50	Male	1	1	1	1	0.90	0.88

Apart from the above-mentioned techniques and line of research in predictive analysis, it was found that predictive analysis was also implemented in finance [11-12] domain, marketing domain [13], Manufacturing [14-15], Retail management [16-18].

As a result, it is clear that there are several methods and strategies for using predictive analysis to solve a variety of issues. All of the approaches that the writers provide have their benefits as well as their drawbacks. Most of the bottlenecks are brought on by a lack of application of benchmarked research. The open research questions in the field of predictive analysis are covered in the following section.

5. Open Research Issues

Recent the power of computation merge its highest peak and availability of data especially healthcare data in plenty created great scope of predictive analysis but it has many research gap.

5.1 Less Innovation in the Methods

Most of current techniques to predictive analysis are somewhat concerned with forecasting future occurrences or behaviours. These models examine a sizable quantity of data with respect to find patterns and trends that they then employ to predict future results. Though only few studies are judged to have made a major contribution to predictive analysis.

5.2. Less focus on complicated cases

Current research methodologies are essentially tested on datasets. Such datasets, however, are not well evaluated and do not accurately demonstrate or forecast.

5.3 Regular Carry Over Limitation

There are a number of restrictions related to, like clearly evaluating prognosis. Because existing strategies concentrate on macro rather than micro concerns, It's discovered that precision is not optimized.

5.4 Fewer Studies with Benchmarks

It has been discovered that there are less research approaches that have been benchmarked. Aside from benchmarking, there are fewer investigations whose results are shown to be inferior to the proper approach.

5.5 Less emphasis on Complexity of Computation

There aren't many examples of computationally efficient algorithms that do predictive analysis on bigger datasets among the papers that are already available.

➢ Most of the existing work has been done on binary classification problems and does not predict the risk of developing comorbidity.

➢ To bridge the gap between the raw Electronic Healthcare Record (EHR) data and the endpoint analytical tasks, like risk prediction or chronic disease.

➢ Existing research use machine learning algorithms based on patient attributes to meet this need; however, they are biased and have high dimensional data issues.

6. Conclusion

The study's findings clearly demonstrate that current research methods have placed less emphasis on new types of dynamic data that are more complicated in terms of structure, heterogeneity, uncertainty, etc. The research community pays relatively little attention to a number of important research issues, such as computing efficiency, data on comorbid disorders, adoption of difficult situations, etc. Predictive analysis studies will need greater focus in order to secure the acceptance of predictive analysis for emerging communication-related technologies. These studies will focus on bigger and comorbid illnesses data. As a result, the focus of our future effort will be on solving these open research problems. We will use data on comorbid conditions to start our inquiry, and we'll also present a method to reduce the intricacy required.

References:

[1] C. Harrison, et al., Comorbidity versus multimorbidity: Why it matters, J. Multimorb. Comorb., 11, 2633556521993993 (2021).

[2] Kingston, et al., Projections of multi-morbidity in the older population in England to 2035: estimates from the Population Ageing and Care Simulation (PACSim) model, Age Ageing 47 (2018) 374–380.

[3] Mohanad M. Alsaleh, et al., Prediction of disease comorbidity using explainable artificial intelligence and machine learning techniques: A systematic review.

[4] E. Ge, Y. Li, S. Wu, E. Candido, X. Wei, Association of pre-existing comorbidities with mortality and disease severity among 167,500 individuals with COVID-19 in Canada: A population-based cohort study, PLoS One 16 (2021) e0258154.

[5] Y.K. Lee, et al., The relationship of comorbidities to mortality and cause of death in patients with differentiated thyroid carcinoma, Sci. Rep. 9 (2019) 11435.

[6] J.F. Figueroa, et al., International comparison of health spending and utilization among people with complex multimorbidity, Health Serv. Res. 56 (Suppl 3) (2021) 1317–1334.

[7] S.I. Cho, S. Yoon, H.-J. Lee, Impact of comorbidity burden on mortality in patients with COVID-19 using the Korean health insurance database, Sci. Rep. 11 (2021) 6375.

[8] D. Sarfati, B. Koczwara, C. Jackson, The impact of comorbidity on cancer and its treatment, CA Cancer J. Clin. 66 (2016) 337–350.

[9] J.F. Piccirillo, I. Costas, The impact of comorbidity on outcomes, ORL J. Otorhinolaryngol. Relat. Spec. 66 (2004) 180–185.

[10] R. Gijsen, et al., Causes and consequences of comorbidity: A review, J. Clin. Epidemiol. 54 (2001) 661–674.

[11] Daniel Broby, The use of predictive analytics in finance, The Journal of Finance and Data Science 8 (2022) 145–161

[12] Suraj Patil, et al., Predictive Modeling For Credit Card Fraud Detection Using Data Analytics. International Conference on Computational Intelligence and Data Science (ICCIDS 2018). DOI : 10.1016/j.procs.2018.05.199

[13] Joe F. Hair Jr, Knowledge creation in marketing: the role of predictive analytics, European Business Review Vol. 19 No. 4, 2007 pp. 303-315 q Emerald Group Publishing Limited 0955-534X DOI 10.1108/09555340710760134

[14] Rameshwar Dubey, et al., Big Data and Predictive Analytics and Manufacturing Performance: Integrating Institutional Theory, Resource-Based View and Big Data Culture, British Journal of Management, Vol. 30, 341–361 (2019) DOI: 10.1111/1467-8551.12355.

[15] David Lechevalier, et. al., Towards a Domain-Specific Framework for Predictive Analytics in Manufacturing, 2014 IEEE International Conference on Big Data, 978-1-4799-5666-1/14/$31.00 ©2014 IEEE

[16] E.T. Bradlow et al., The Role of Big Data and Predictive Analytics in Retailing, Journal of Retailing 93 (1, 2017) 79–95.

[17] Disha Budale et al., Predictive Analytics in Retail Banking, International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249 – 8958, Volume-2, Issue-5, June 2013.

[18] Chejarla Venkat Narayana et al., Machine Learning Techniques To Predict The Price Of Used Cars Predictive Analytics in Retail Business, Proceedings of the Second International Conference on Electronics and Sustainable Communication Systems (ICESC-2021) IEEE Xplore Part Number: CFP21V66-ART; ISBN: 978-1-6654-2867-5.

A. Kline, et al. Multimodal Machine Learning in Precision Health. arXiv [cs.LG] (2022).

[19] T. Linden, et al., An Explainable Multimodal Neural Network Architecture for Predicting Epilepsy Comorbidities Based on Administrative Claims Data, Front. Artif. Intell. 4 (2021) 610197.

[20] England, N. H. S. Improving outcomes through personalised medicine. NHS England https://www.england.nhs.uk/wp-content/uploads/2016/09/improvin g-outcomes-personalised-medicine.pdf (2016).

[21] J. Zhao, et al., Learning from Longitudinal Data in Electronic Health Record and Genetic Data to Improve Cardiovascular Event Prediction, Sci. Rep. 9 (2019) 717.

[22] J. Deng, T. Hartung, E. Capobianco, J.Y. Chen, F. Emmert-Streib, Editorial: Artificial Intelligence for Precision Medicine, Front. Artif. Intell. 4 (2021) 834645.

[23] P. Akram, L. Liao, Prediction of comorbid diseases using weighted geometric embedding of human interactome, BMC Med. Genomics 12 (2019) 161.

[24] R.F. Wolff, et al., PROBAST: A Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies, Ann. Intern. Med. 170 (2019) 51–58.