International Journal of Computer Science Engineering Techniques (IJCSE) Volume 10 Issue 3 , May-June - 2026

ISSN : 2455-135X

A Review of Human Disease Prediction Models Using Deep Learning and Symptom Analysis

Sujata Ramesh Ambhore [1], Shital Nivrutti Katkade[2]. Reema Ashok Lahane[3], Ramesh Raybhan Manza[4]

[1]Dr. Babasaheb Ambedkar Marathwada University, Chhatrapati Sambhajinagar, Maharashtra.
[2]Dr. Babasaheb Ambedkar Marathwada University, Chhatrapati Sambhajinagar, Maharashtra.
3Dr. Babasaheb Ambedkar Marathwada University, Chhatrapati Sambhajinagar, Maharashtra.
4Dr. Babasaheb Ambedkar Marathwada University, Chhatrapati Sambhajinagar, Maharashtra.

[1]ambhoresujata2@gmail.com

Abstract - *Disease diagnosis is difficult in the modern world since hospital visits are sometimes expensive and time-consuming, especially for people who live far from medical facilities. The Disease Predictor provides a practical and affordable solution by estimating the likelihood of a disease based on user-input symptoms using deep learning and symptom analysis. With an emphasis on deep learning methods such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Graph Neural Networks (GNNs), this paper investigates predictive modeling for the prediction of human diseases. The suggested methodology improves the efficiency and accuracy of diagnosis by assessing symptom-based inputs. By filling a research gap in the integration of multimodal data for better prediction, this study advances automated, scalable healthcare systems that put patient accessibility and early diagnosis first.*

Keywords - *Artificial Neural Network (ANN), Electronic Health Record (EHR), Convolutional Neural Network (CNN), Graph Neural Network (GNN), Cardiovascular Disease (CVD)*

Graphical Abstract - *The graphical abstract showcases the workflow for human disease prediction which comprises the following: Symptom Input, Data Preprocessing, Feature Extraction, Deep Learning Model, Disease Prediction Output*

1. Introduction

Since going to the hospital is costly and time-consuming in today's world, not everyone can afford it. The user may also find it difficult if they reside far from hospitals and medical specialists because the problem cannot be detected. The aforementioned procedure might therefore be carried out using automated software that saves time and money, which might benefit the patient and make the process go more easily. The user can use the Disease Predictor to determine the likelihood of a particular disease based on its symptoms. illness could have few options. Thus, this system may be advantageous to individuals. The "Disease Prediction" approach, which relies on predictive modeling, predicts the user's condition based on the symptoms they input. The method evaluates the symptoms entered by the user and returns the condition's likelihood The science of creating computer systems that can learn The science of creating computer systems that can learn from experience and data is known as machine learning. Deep learning is a specific part of this larger field. Consequently, the model's training and testing protocols must be adhered to. Human diseases are a broad category of medical problems that conflict with the body's natural functioning and frequently result in suffering, either physical or psychological. Numerous variables, such as genetic predisposition, environmental effects, lifestyle choices, infections, or even unidentified reasons, might contribute to the development of these disorders. In addition to affecting people, diseases often place a strain on families, healthcare systems, and entire societies. This review's objective is to compile and evaluate deep learning and machine learning methods for predicting human diseases in order to determine their present strengths, weaknesses, and suitability for use in actual healthcare environments. The ultimate objective is to give researchers and practitioners a clear road map of current approaches while pointing out areas that need more attention to make these models scalable and clinically effective.

ISSN : 2455-135X

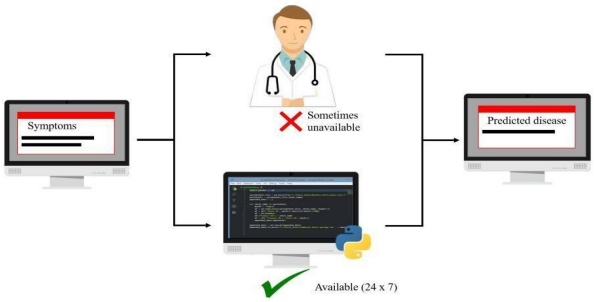

Fig1: human disease prediction based on their symptoms

1.1 Problem Statement

In spite of the current development of artificial intelligence and deep learning, the correct and easy diagnosis of disease is still a big challenge in the current healthcare system. The traditional diagnosis methods involve making a visit to the hospital, expert advice and expensive medical testing – a process that is resource consuming and may not be possible for those who live in remote or resource limited areas. Current disease prediction models are mostly based on medical imaging data or structured clinical records, and symptom-based disease prediction models still have challenges, including low interpretability, weak generalization ability, and failure to consider the complex relationships between symptoms and diseases.

Moreover, many studies have analyzed various models, but have not determined the "best" architecture for practical healthcare applications. Hence, it is essential to have an intelligent disease prediction framework, which analyzes the symptom data with advanced deep learning methods that are able to model non-linear and relational dependency and enhance the prediction's accuracy and early diagnosis.

1.2 Objectives

The main aims of the present study are:

1. To review previous machine learning and deep learning techniques for human disease prediction.
2. To explore disease prediction based on symptom using deep learning methods.
3. To develop an effective prediction model based on advanced neural networks architectures.
4. For assessing the performance of the models on conventional evaluation metrics including Accuracy, Precision, Recall, F1-Score and ROC–AUC.
5. To identify the most suitable model for real-world healthcare applications.
6. For early detection of disease and to aid in intelligent healthcare decision systems.

1.3 Organization of the article

The rest of this article is structured as follows:

1. The Literature Review is included in Section 2.
2. Materials and Methods are discussed in Section 3.
3. Results and Discussion is covered in Section 4.
4. The study is summarized in Section 5 and future research directions are pointed out.

1.4 Novelty and Contribution of the Present Study

The present study brings a novelty to the field of intelligent healthcare systems in regard to the following aspects Analyzes and presents a complete review of deep learning models such as CNN, RNN, and Graph Neural Networks for disease prediction.Targets specifically a disease prediction system based on symptoms, making it more accessible than imaging-dependent systems. Emphasizes the significance of Graph Neural Networks (GNN) in the representation of relationships between symptoms and diseases. Combines data preprocessing, feature selection and deep learning classification into a single predictive model. Provides comparative analysis to determine the most successful model for diagnosing disease. Solves practical problems in healthcare including scalability, interpretability and early diagnosis support. Offers insights for future research on explainable AI and multimodal healthcare analytics.

2. Literature Review

The rise of deep learning models for predicting human diseases has attracted considerable interest in recent years. Researchers have utilized a variety of deep learning architectures, including convolutional neural networks (CNNs), recurrent neural networks (RNNs), along with transformer-based models, to analyze intricate medical data for disease prediction[1]. These approaches have shown impressive capabilities in managing large datasets, extracting pertinent features, and enhancing diagnostic accuracy. One of the first uses of deep learning in disease prediction was the application of CNNs in medical imaging. For example, Esteva et al. (2017) employed CNNs to classify skin lesions, achieving performance levels comparable to that of dermatologists[2]. Likewise, Rajpurkar et al. (2018) used a deep learning algorithm to detect pneumonia from chest X-rays, demonstrating its ability to surpass radiologists in certain instances. Beyond imaging, symptom-based analysis has been investigated by researchers to improve prediction accuracy[3]. To predict patient outcomes using electronic health records (EHRs), Choi et al. (2016) presented the "RETAIN" model, which made advantage of attention mechanisms in RNNs.(EHRs).

Natural language processing (NLP) methods have also been used in symptom-based models, such those by Nguyen et al. (2020), to examine clinical notes and patient-reported symptoms, showing promise for incorporating subjective data into prediction algorithms[4]. The integration of several kinds of data, including clinical, genomic, along with lifestyle Details,

ISSN : 2455-135X

is becoming a promising approach[5]. Hybrid models that combine CNNs with graph neural networks (GNNs) have demonstrated potential in analyzing diverse data types[6]. Nevertheless, important issues related to ethics, patient privacy, and scalability still need to be addressed[7][8].(et. Al Sumit Sharma 2020) Heart disease prediction has been a critical area of research in healthcare, leveraging machine learning and deep learning techniques to enhance diagnostic accuracy[8]. Studies have employed algorithms like Logistic Regression, KNN, SVM, Naïve Bayes, and Random Forest, each offering varying levels of performance[9]. Logistic Regression and KNN have shown efficacy in handling linear and distance-based data patterns, while SVM and Naïve Bayes have been effective for high- dimensional and probabilistic models, respectively[9]. Recent advancements have emphasized hyperparameter optimization to improve predictive accuracy. Random Forest, due to its ensemble learning capability, has consistently outperformed others in terms of precision and robustness[9]. This research builds on these foundations, utilizing these models with optimization techniques, and identifies Random Forest as the most accurate predictor for heart disease detection. In order to predict common human diseases, (et.al Jabir Al Nahian 2022) combines data mining and machine learning techniques. The study uses four classifiers: Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Logistic Regression, and Random Forest. The Random Forest algorithm showed the highest accuracy in disease prediction, demonstrating its efficacy and robustness[10]. The study emphasizes the potential of using survey data and sophisticated algorithms for early disease detection, which contributes to better healthcare decision-making and analytics. In the proposed research in the biomedical field, machine learning (ML) and pattern recognition have the capability to increase the precision of disease identification and diagnostic techniques. Additionally, they support the impartiality of the decision-making procedure. A dependable method for creating enhanced, automated algorithms to evaluate multi-modal, high-dimensional biological data is machine learning (ML). A comparison of several machine learning algorithms for the diagnosis of various illnesses, such as diabetes and heart disease, is provided in this survey study[11][12]. It highlights the variety of machine learning techniques and algorithms used in decision- making and illness diagnosis. In this instance, Naïve Bayes achieved the highest accuracy. A significant worldwide health concern, cardiovascular disease (CVD) requires precise predictive models for early detection and prevention[13].Various machine learning algorithms have been explored for this purpose, including Support Vector Machine (SVM), Decision Tree, and K-Nearest Neighbors (KNN). While SVM and Decision Tree demonstrated strong predictive capabilities, KNN showed comparatively lower performance[14]. Among all models, the Artificial Neural Network (ANN) binary classification model achieved the highest accuracy, highlighting its effectiveness in capturing complex patterns and nonlinear relationships in medical data. These findings emphasize the potential of ANN in improving cardiovascular disease prediction and diagnosis[14][15]. The body of research suggests that both traditional machine learning and contemporary deep learning techniques make substantial contributions to the study of disease prediction. Classical models are still useful for their interpretability and effectiveness in some situations, even though deep learning techniques are excellent at identifying intricate patterns and combining multimodal data. A well-rounded strategy that incorporates both contemporary and conventional methods seems to hold promise for enhancing precision and usability in medical applications.

ISSN : 2455-135X

Sr. No	Study	Technique/Model	Dataset	Disease Focus	Accuracy	Limitations
1.	Esteva et al. (2017)	CNN	ISIC Skin Cancer Dataset	Skin Cancer	91%	Limited to image-based diagnosis only.
2.	Rajpurkar et al. (2018)	CNN	ChestX-ray14	Pneumonia	92%	Relies heavily on imaging data.
3.	Choi et al. (2016)	RETAIN (RNN with Attention)	MIMIC-III	General Predictions	80%	Limited interpretability of results.
4.	Nguyen et al. (2020)	NLP-based DL Models	Clinical Notes Dataset	Symptom Analysis	85%	Requires high-quality textual data.
5.	Lundberg & Lee (2017)	Explainable AI (SHAP)	Various	General Predictions	88%	Computationally intensive for large datasets.
6.	Miotto et al. (2016)	Deep Autoencoders	EHR Data	Multi-disease Prediction	84%	Lacks integration of genomic/lifestyle data.
7.	Lee et al.(2020)	Hybrid Model (CNN+ GNN)	Genomic + Imaging Data	Cancer	93%	Combined imaging and genomic data to improve cancer prediction accuracy.
8.	Sumit et al.(2020)	Logistic regression, KNN, SVM Naïve Bayes, Random Forest, Hyperparameter Optimization	Heart Disease Dataset	Heart Disease	89%	Limited exploration of deep learning-specific neural architectures.
9.	(Jabir Al Nahian et al., 2022)	Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Logistic Regression, Random Forest	Survey-based dataset	Common Human Diseases	87%	Dependence on survey data, which may introduce bias or inconsistencies. Limited focus on deep learning
10.	(P. Hamsaga yathri et al.2021)	Naïve Bayes, SVM, Decision Tree, KNN	Symptom-based dataset	Diabetes, Heart Disease	82%	Naïve Bayes: Assumes feature independence. SVM: High computational complexity. Decision Trees: Prone to overfitting. KNN: Struggles with high-dimensional data.
11.	Sayed Pasha, P et al. (2020)	SVM, KNN, Decision Tree, ANN (Binary Model)	Heart Disease UCI	Cardiovascular Disease	90%	KNN showed lower performance; Decision Tree prone to overfitting; ANN requires high computation
12.	Smith et al. (2021)	Transformer-based NLP	Clinical Text Data	Cancer Diagnosis	86%	Requires large labeled datasets.
13.	Li et al. (2019)	GANs	Medical Imaging Dataset	Tumor Detection	89%	Needs careful training to avoid mode collapse.
14.	Wang et al. (2021)	Graph Neural Networks (GNN)	Genomic Data	Genetic Disorder Prediction	91%	High computational complexity
15.	Kim et al. (2022)	Federated Learning	Distributed Hospital Data	COVID-19 Prediction	90%	Requires reliable communication channels.

First-Order Heading (SIZE 12 &BOLD)

2.1. Second-Order Heading (SIZE 10 & BOLD & Italic)

2.1.1. Third-Order Heading (Size 10 & Italic)

Fourth-Order Heading (Text, no numbering) (*Size 10 & Italic)*

3. Materials and Methods (SIZE 12 &BOLD)

The materials and methods section should contain sufficient detail so that all procedures can be repeated. It may be divided into headed subsections if several methods are described. (Size 10 & Normal)

4. Results and Discussion (SIZE 12 &BOLD)

ISSN : 2455-135X

4.1. Subheadings (Size 10 & bold & Italic)

The results and discussion may be presented separately, or in one combined section, and may optionally be divided into headed subsections. (Size 10 & Normal)

((Main Text Paragraphs. Please make the first reference to a display item bold **(Figure 1)**. Do not abbreviate Figure, Equation, etc.; display items are always singular, i.e., Figure 1 and 2. Equations are always singular, i.e., Equation 1 and 2, and should be inserted using the Equation Editor, not as graphics, in the main text. Display items and captions should be inserted in-line within the main text)). ((References should be in the parenthesis and appear after punctuation. [1,2] If you have used reference management software such as EndNote to prepare your manuscript, please convert the fields to plain text by selecting all text with [ctrl]+[A], then [ctrl]+[shift]+[F9]). [3–5] Footnotes should not be used in the text. Instead, additional information can be added to the Reference list.

ISSN : 2455-135X

Your paper must use a **page size corresponding to Letter** which is 8.5" wide and 11" long. The margins must be set as follows:

- Top = 1"
- Bottom = 1"
- Left = Right = 0.7"
- Header = Footer = 0.5"

The two-column format in the manuscript must be with a space of 0.25" between columns. The entire document should be in **Times New Roman Font**. Type other font types may be used if needed for special purposes. Recommended font sizes are shown in Table 2.

Table 1. Title of the table (Size 8 & Bold for the Table Caption)
((Table captions should be placed above the tables.))

(SIZE 10 & Bold)	Table Header	Table Header
Heading	(SIZE 10)	Values
Heading	Values	Values
Heading	Values	Values

Small table or figure must be placed in the double column and positioned either at the top or at the bottom of the page.

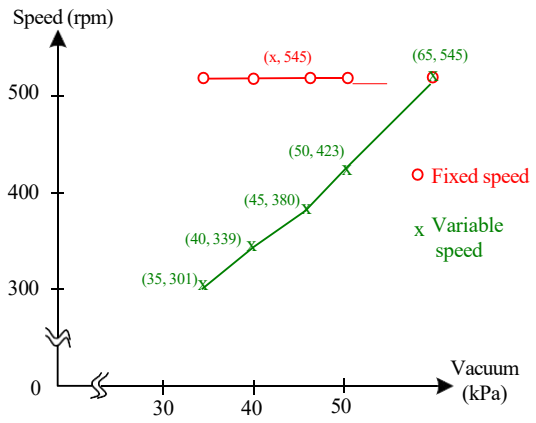

Fig. 1 Example for small figure (Size 8 & Bold for the Fig Caption)
((Figure Caption need to be placed below the figures. Note: Please do not combine figure and caption in a textbox or frame.))

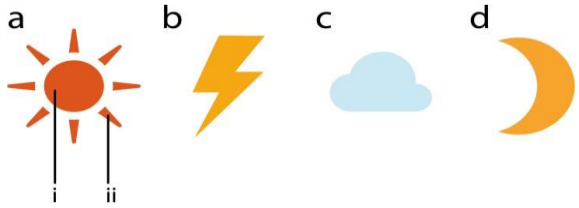

Fig. 2 Example for figure consists of multiple panels: Representations of some common weather symbols. (a) The Sun with (i) Core, (ii) Rays, (b) Thunder bolt, (c) Cloud, and (d) Moon.

Table 2. Recommended font sizes (Size 8 & Bold for the Table Caption)
Note: Please do not combine table and caption in a textbox or frame and do not submit tables as graphics, please use Word's "insert table" function.

Heading level (SIZE 10 & Bold)	Example	Font size and style
Title of the Paper (centered)	Title of Research Article	22-point, normal
1st-level heading	**1. Introduction**	12-point, bold
2nd-level heading	***1.1. Printing Area***	10-point, bold, Italic
3rd-level heading	*1.1.1. Run-in Heading*	10-point, italic
4th-level heading	*Lowest Level Heading*	10-point, italic, no numbering
Figure and Table	**Table 1. Caption follows** **Fig. 1 Caption follows**	8-point, bold
References	[1] S. Zhang, C. Zhu, J. K. O. Sin, and P. K. T. Mok, "A Novel Ultrathin Elevated Channel Low-Temperature Poly-Si TFT," *IEEE Electron Device Letters*, vol. 20, no. 4, pp. 569–571, 1999.	9-point, normal

[a])((Table Footnote)); [b]) …(SIZE 8)
Source: Text follows (SIZE 8 Italic)

ISSN : 2455-135X

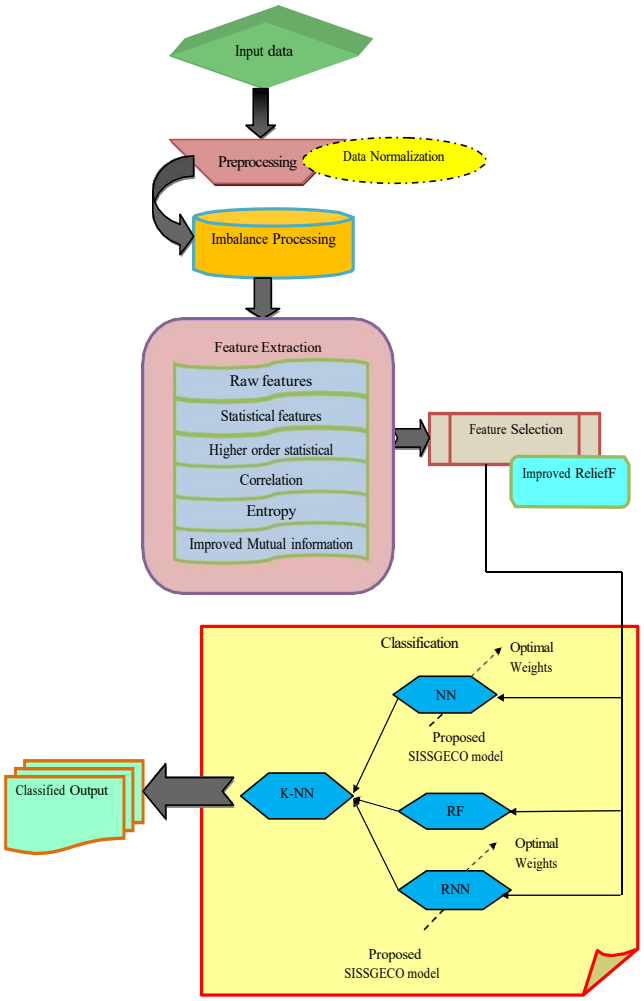

Fig. 3 Example for large figures (Size 8 & Bold for the Figure Caption)
[Large Figure / tables may span across both columns and must be positioned either at the top or at the bottom of the page]

Figure must be positioned either at the top or at the bottom of the page. They should be referred to as (see Figure 1, etc.) in the main text. Please check all figures in your paper both on screen and on a black-and-white hardcopy. When you check your paper on a black-and-white hardcopy, please ensure that:

- The colors used in each figure contrast well,
- The image used in each figure is clear,
- All text labels in each figure are legible, &
- Reproduced with permission. [Ref.] Copyright Year, Publisher.

((Permission statement required for all figures reproduced or adapted from previously published articles/sources. CC-BY content can be used without asking permission, but the source must be attributed: Reproduced under terms of the CC-BY license. [ref] Copyright Year, The Authors, published by [Publisher]. However, permission must be obtained for reproduction of content published under CC-BY-NC/ND/SA licenses; delete if not applicable.))

If a figure consists of multiple panels, they should be ordered logically and labelled with lower case roman letters (i.e., a, b, c, etc.). If it is necessary to mark individual features within a panel, this may be done with lowercase Roman numerals, i, ii, iii, iv, etc. All labels should be explained in the caption. Panels should not be contained within boxes unless strictly necessary. (See Figure 2)

Equations

Equations and formulae should be typed in equation plug-in like Microsoft Word's equation tool, and numbered consecutively with Arabic numerals in parentheses on the right-hand side of the page (if referred to explicitly in the text). They should also be separated from the surrounding text by one space. They should be referred to as Equation 1, etc. in the main text.

$$\rho = \frac{E}{J_c\left(T = \text{const.}\right) \cdot \left(P \cdot \left(\frac{\vec{E}}{E_C} \right)^m + \left(1 - P\right) \right)} \qquad (1)$$

ISSN : 2455-135X

Displayed equations are centered and set on a separate line. Please avoid rasterized images for equations, tables, flow charts, algorithms, datasets, line-art diagrams and schemas. Whenever possible, use vector graphics instead.

5. Conclusion (SIZE 12 &BOLD)

The Conclusions section should clearly explain the main findings and implications of the work, highlighting its importance and relevance. (SIZE 10)

Conflicts of Interest (SIZE 12 &BOLD)

This section is compulsory. A competing interest exists when professional judgment concerning the validity of research is influenced by a secondary interest, such as financial gain. We require that our authors reveal any possible conflict of interest in their submitted manuscripts. If there is no conflict of interest, authors should state that "The author(s) declare(s) that there is no conflict of interest regarding the publication of this paper." (SIZE 10)

Funding Statement (SIZE 12 &BOLD)

Authors should state how the research and publication of their article was funded, by naming financially supporting bodies followed by any associated grant numbers in square brackets. (SIZE 10)

Acknowledgments (SIZE 12 &BOLD)

An Acknowledgements section is optional and may recognise those individuals who provided help during the research and preparation of the manuscript. Other references to the title/authors can also appear here, such as "Author 1 and Author 2 contributed equally to this work. (SIZE 10)

References (FONT SIZE 12 & BOLD)

Note: Authors should ensure that their citations are accurate; Authors should not cite sources that they have not read; & Authors should not preferentially cite their own or their friends', peers', or institution's publications.

(Authors are requested to follow the below reference format in a strict manner)

Examples of reference items of different categories shown in the References section include:

- example of a journal article in [1]
- example of a conference paper in [2]
- example of a book in [3]
- example of a book in a series in [4]
- example of a patent in [5]
- example of a website in [6]
- example of a web page in [7]
- example of a datasheet in [8]
- example of a master's thesis in [9]
- example of a technical report in [10]
- example of a data book as a manual in [8]

[1] (Font Size 9, Line Spacing 1.15) S. Zhang, C. Zhu, J. K. O. Sin, and P. K. T. Mok, "A Novel Ultrathin Elevated Channel Low-Temperature Poly-Si TFT," *IEEE Electron Device Letters*, vol. 20, no. 4, pp. 569–571, 1999. [CrossRef] [Google Scholar] [Publisher Link]

[2] Yi W., Jongwoo L., and Ming-Hsuan Y., "Online Object Tracking: A Benchmark," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2411-2418, 2013. [CrossRef] [Google Scholar] [Publisher Link]

[3] S. M. Metev, and V. P. Veiko, *Laser Assisted Microtechnology*, 2nd ed., R. M. Osgood, Jr., Ed. Berlin, Germany: Springer-Verlag, vol. 5, no. 3, pp. 300-320, 1998. [CrossRef] [Google Scholar] [Publisher Link]

[4] J. Breckling, Ed., *The Analysis of Directional Time Series: Applications to Wind Speed and Direction*, ser. Lecture Notes in Statistics. Berlin, Germany: Springer, vol. 61, no. 1, pp. 200-220, 1989. [CrossRef] [Google Scholar] [Publisher Link]

[5] R. E. Sorace, V. S. Reinhardt, and S. A. Vaughn, "High-Speed Digital-to-RF Converter," *U.S. Patent 5668842*, vol. 20, no. 2, pp. 300-325, 1997.

[6] The IEEE Website, 2002. [Online]. Available: http://www.ieee.org/

[7] FLEX Chip Signal Processor (MC68175/D), *Motorola*, vol. 15, no. 3, pp. 250-275, 1996.

[8] *PDCA12-70 Data Sheet*, OptoSpeedSA, Mezzovico, Switzerland.

[9] A. Karnik, "*Performance of TCP Congestion Control with Rate Feedback: TCP/ABR and Rate-Adaptive TCP/IP*," M.E. Thesis, Indian Institute of Science, Bangalore, India, 1999.

[10] J. Padhye, V. Firoiu, and D. Towsley, "*A Stochastic Model of TCP Reno Congestion Avoidance and Control*," University of Massachusetts, Amherst, MA, CMPSCI Technical Report, 1999.

[11] *Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specification*, IEEE Std., vol. 12, no. 11, pp. 260-280, 1997.

Appendix 1 etc. (SIZE 12 &BOLD)

Appendices, if present, must be marked 1, 2, 3.