

Artificial intelligence for the re-modelling of healthcare system

A DISSERTATION

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE AWARD OF THE DEGREE

OF

MASTER OF SCIENCE

In

BIOTECHNOLOGY



**DEPARTMENT OF BIOTECHNOLOGY
DELHI TECHNOLOGICAL
UNIVERSITY (Formerly Delhi
College of Engineering) Bawana
Road, Delhi - 110042**

Submitted By
Ankit
2K20/MSCBIO/42

Under the supervision of:
Prof. Yasha Hasija
Department of Biotechnology
Delhi Technological University

DEPARTMENT OF BIOTECHNOLOGY
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi - 110042

CANDIDATE'S DECLARATION

I Ankit, Roll Number: 2K20/MSCBIO/42, student of M.Sc. Biotechnology, hereby declare that the work which is presented in the Major Project entitled — “**Artificial intelligence for the re-modelling of healthcare system**” in the fulfillment of the requirement for the award of the degree of Master of Science in Biotechnology and submitted to the Department of Biotechnology, Delhi Technological University, Delhi, is an authentic record of my own carried out during the period from January- May 2022, under the supervision of Prof. Yasha Hasija.

The matter presented in this report has not been submitted by me for the award for any other degree of this or any other Institute/University. The work has been accepted in SCI/SCI expanded /SSCI/Scopus Indexed Journal OR peer reviewed Scopus Index Conference with the following details:

Title of the Paper: Artificial Intelligence and Digital Pathology Synergy: For detailed, accurate and predictive analysis of WSIs

Author Names: Ankit and Yasha Hasija

Name of Conference: 2022 8th International Conference on Advanced Computing and Communication Systems (ICACCS).

Conference Date and Venue: 25th-26th March, Sri Eshwar College of Engineering, Coimbatore, Tamilnadu, India.

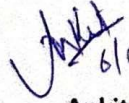
Registration: Done

Status of Paper: Acceptance Received

Date of Paper Communication: 27th February 2022

Date of Paper Acceptance: 3rd March 2022

Date of Paper Publication: NA


6/05/22
Ankit
2K20/MSCBIO/42

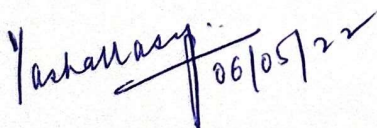
**DEPARTMENT OF BIOTECHNOLOGY
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi - 110042**

SUPERVISOR CERTIFICATE

I hereby certify that the Project dissertation titled “**Artificial intelligence for the re-modelling of healthcare system**” which is submitted by **Ankit, 2K20/MSCBIO/42**, Department of Biotechnology, Delhi Technological University, Delhi in partial fulfillment of the requirement for the award of the degree of Master of Science, is a record for the project work carried out by the student under my supervision. To the best of my knowledge this work has not been submitted in part or full for any Degree or Diploma to this University or elsewhere.

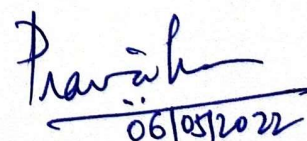
Place: Delhi

Date: 6 May 2022



**Prof. Yahsa Hasija
(Supervisor)**

Department of Biotechnology
Delhi Technological University



**Prof. Pravir Kumar
Head of Department
Department of Biotechnology
Delhi Technological University**

Acknowledgement

I would like to express my gratitude towards my supervisor, Prof. Yasha Hasija, for giving me the opportunity to do research and providing invaluable guidance throughout this research. Her dynamism, vision, sincerity, and motivation have deeply inspired me. She has motivated me to carry out the research and to present my work as clearly as possible. It was a great privilege and honour to work and study under her guidance. I am extremely grateful for what she has offered me. Her insightful feedback pushed me to sharpen my thinking and brought my work to a higher level. I am equally grateful and wish to express my wholehearted thanks to respected Mr Rajkumar sir for their kind support and help in the course of my research work. I am extremely grateful to my parents for their love, prayers, caring and sacrifices for educating and preparing me for my future. I felt fortunate to be able to study in such an excellent institution "Delhi Technological University", I would like to thank all the DTU authorities for giving me the opportunities to study in such a reputed university. Finally, my thanks go to all the people who have supported me to complete the research work directly or indirectly.

Ankit
06/05/22

Ankit
2K20/MSCBIO/42

ABSTRACT

Artificial intelligence had re-modelled the various fields and made ease for various difficult tasks which were hard to do by traditional approaches in the related fields. AI impacted entertainment, IT, mechanical and various type of industries and changes lifestyle of the world. Healthcare sector is not left behind of this there are various healthcare sectors in which AI is playing crucial role and providing support to healthcare infrastructure. AI can improve the workflow of various diagnostic and treatment facilities. AI can analyse images and this feature has robust application in pathological tasks and now pathologists can visualize histopathological images directly on the computer screen. AI can be used in drug discovery and designing as they can predict how the drug will going to react to a particular molecule. AI has still to do more in healthcare sector and this thesis is done in order to provide evidence of the positive results of these two fields synergy.

Contents

Artificial intelligence for the re-modelling of healthcare system.....	10
1.1 WHAT IS ARTIFICIAL INTELLIGENCE.....	10
1.1.1 Types of AI.....	10
2 Machine learning.....	10
2.1 Supervised Learning:-.....	11
2.1.1 Discission Tree.....	11
2.1.2 Naïve Bayes:	11
2.1.3 Support Vector Machine:	12
2.2 Unsupervised Learning	12
2.2.1 K-means clustering	13
2.3 Semi-supervised Learning	13
3 Deep Neural Networks	15
3.1.1 Convolutional Neural Network (CNN)	15
3.1.2 Fully Convolutional Neural Network (FCN)	16
3.1.3 Generative adversarial network (GAN)	17
3.1.4 Recurrent Neural Network (RNN)	17
4 AI Publications Boom In Modern Era	19
5 Application of AI synergy in Healthcare	21
5.1 Application of AI in Pathology:-.....	21
5.1.1 How Does It works.....	22
5.2 In Bioinformatics:	22
5.3 In Medical Imaging Technology.....	23
5.4 In pharmaceuticals.....	23
6 Drug Discovery and Regression Models Accuracy Analysis Using Machine Learning Algorithms.....	23
7 Literature Review.....	24
8 Methodology	25
8.1 Biological Data Extraction using ChEMBL Database:.....	26
8.2 Retrieve Bioactivity data reported as IC50 values in nM (nanomolar) unit:	27
8.3 Remove Duplicate Data and Missing data:	28
8.4 Compound Labelling :.....	28
8.5 Pre-Processing of Datasets.....	29
8.5.1 Removal of unwanted features from datasets:	29

8.5.2	Lipinski rule categorization to understand relationships:.....	29
8.5.3	Exploratory Data Analysis:.....	31
8.5.4	Final dataset obtained after pre processing	31
8.6	IC50 to pIC50 conversion:	32
8.7	Conversion of structural and chemical properties into numerical values:.....	32
8.8	QSAR MODEL VALIDATION USING WEKA	33
9	Result and Interpretation of EDA and Statistical Analysis	34
10	Comparison of various regression model	38
11	Conclusion And Future Aspects.....	41
12	References	43

List Of Figures

Figure 1 Overview of different type of AI approaches and various algorithm

categories.14

Figure 2 AI publication curve..... 20

Figure 3 Methodology of thesis..... 26

Figure 4 Retrieving bioactivity data from ChEMBL..... 27

Figure 5 Pie chart showing major subset of IC50 28

Figure 6 Dataset after removal of unwanted features..... 29

Figure 7 Final dataset obtained after pre-processing..... 31

Figure 8 PaDEL- Descriptor software 33

Figure 9 Scatter plot of MW vs Log P of 2 bioactivity classes..... 34

Figure 10 Box plot between MW of 2 bioactivity class..... 35

Figure 11 Box plot between Log P of 2 bioactivity class..... 36

Figure 12 Box plot between pIC50 value of 2 bioactivity class..... 36

Figure 13 Box pot between Number of hydrogen acceptor of 2 bioactivity class 37

Figure 14 Box plot between number of hydrogen donor of 2 bioactivity class 37

Figure 15 waikato environment for knowledge Analysis WEKA..... 38

Figure 16 Comparison of Various regression models Vs R-squared 40

Figure 17 Comparison of Various regression model vs RMSE 40

List of Tables

TABLE I VARIOUS ADVANTAGES AND DISADVANTAGES OF DNN MODELS.

TABLE II TIMELINE SHOWING THE NUMBER OF AI PUBLICATIONS AND ITS CONTRIBUTION PERCENTAGE IN THE AI-RELATED HEALTHCARE PUBLICATIONS.

TABLE III DATASET VALUES AFTER FILTERING DATA ON THE BASIS OF LIPINKSI'S RULE.

TABLE IV Showing various regression models accuracy on the basis of parameters R-squared, Adjusted-R Square, and RMSD.

Artificial intelligence for the re-modelling of healthcare system

1.1 WHAT IS ARTIFICIAL INTELLIGENCE

Artificial intelligence is a computer science branch in which machines are empowered with smartness by using various algorithms and these tasks then easily and smartly performed by those machines.

1.1.1 Types of AI

Artificial intelligence is generally classified into two:-

1.1.1.1 General AI

General AI is a system that learns arbitrary tasks independently and is similar to human judgement

1.1.1.2 Narrow AI

Narrow AI is a system that is built for single task. Narrow AI is further classified into two.

- a) Expert System: This system depends on foundations, rules and knowledge that is been created by humans
- b) Machine Learning: It analyses data and learn directly from features like pattern and relationships.

2 Machine learning

Machine learning is most widely adapted AI technology and is used for learning of various models. Its algorithms improves automatically by learning from the data. Machine learning are categorized into 3 major groups :- supervised, unsupervised, semi-supervised.[10]

2.1 Supervised Learning:-

In supervised learning the input dataset is isolated from the training and testing datasets.[12] The train dataset's output variable is one that has to be predicted or categorised. The algorithms learn patterns from the training dataset and apply them to the test dataset while solving prediction or classification issues.

Mostly used supervised learning are:-

2.1.1 Discission Tree

Trees that aggregate characteristics by ranking them according to their values are known as decision trees. The decision tree is most commonly used to solve categorization issues.[13] Each tree is made up of nodes and branches. Each node represents attributes in a group that need to be categorised, and each branch represents a value that the node can take. Decision trees are a trustworthy and effective decision-making tool that combines high classification accuracy with a simple representation of acquired knowledge to produce reliable and effective results. They've been used in a wide range of medical decision-making scenarios.[32]

2.1.2 Naïve Bayes:

Nave Bayes' primary goal is to work in the text categorization sector. It is mostly used for clustering and classification purposes.[32] The core architecture of Nave Bayes uses conditional probability. It builds trees based on the chance of their occurrence. These trees are also known as Bayesian Networks. The Bayes theorem is applied to features with high

(naive) independence assumptions in naive Bayes classifiers, which are probability-based classifiers. The model is simple to construct and does not require iterative parameter estimation, making it particularly useful in the medical industry. Given the probability distribution, the Bayes classifier can clearly attain the optimal outcome.[32] The Heart Disease Prediction System has also employed the Naive Bayesian Classification technique to build decision assistance. By preserving and digitalizing millions of patients' treatment information, data mining techniques may aid in solving a variety of significant and critical challenges related to health care. The Nave Bayes classification is the best decision assistance system.

2.1.3 Support Vector Machine:

It's the popular ML approach that's largely used for categorisation. The concept of calculating margins underpins SVM. It is mostly used to create margins between distinct classes. The margins are adjusted to maximise the distance between the margin and the classes, lowering classification error.

2.2 Unsupervised Learning

Using unsupervised learning approaches, the data is just used to train a few features. It recognises the data's class using previously learned features when fresh data is introduced. It's typically used for clustering and feature reduction.

2.2.1 K-means clustering

With K as the group number, this algorithm aids in data grouping. Each data point is iteratively assigned to a group based on the features provided. The feature's similarity is then utilised to cluster the data points. The K's centroids are K-means clustering produces clusters and labels for the data.

2.3 Semi-supervised Learning

The advantages of both supervised and unsupervised learning techniques are combined in semi-supervised learning algorithms. If there is existing unlabeled data and gathering the labelled data is a time-consuming task, it might be valuable in domains like machine learning and deep learning. Semi-supervised learning is widely employed in the field of medical science for medical picture classification.

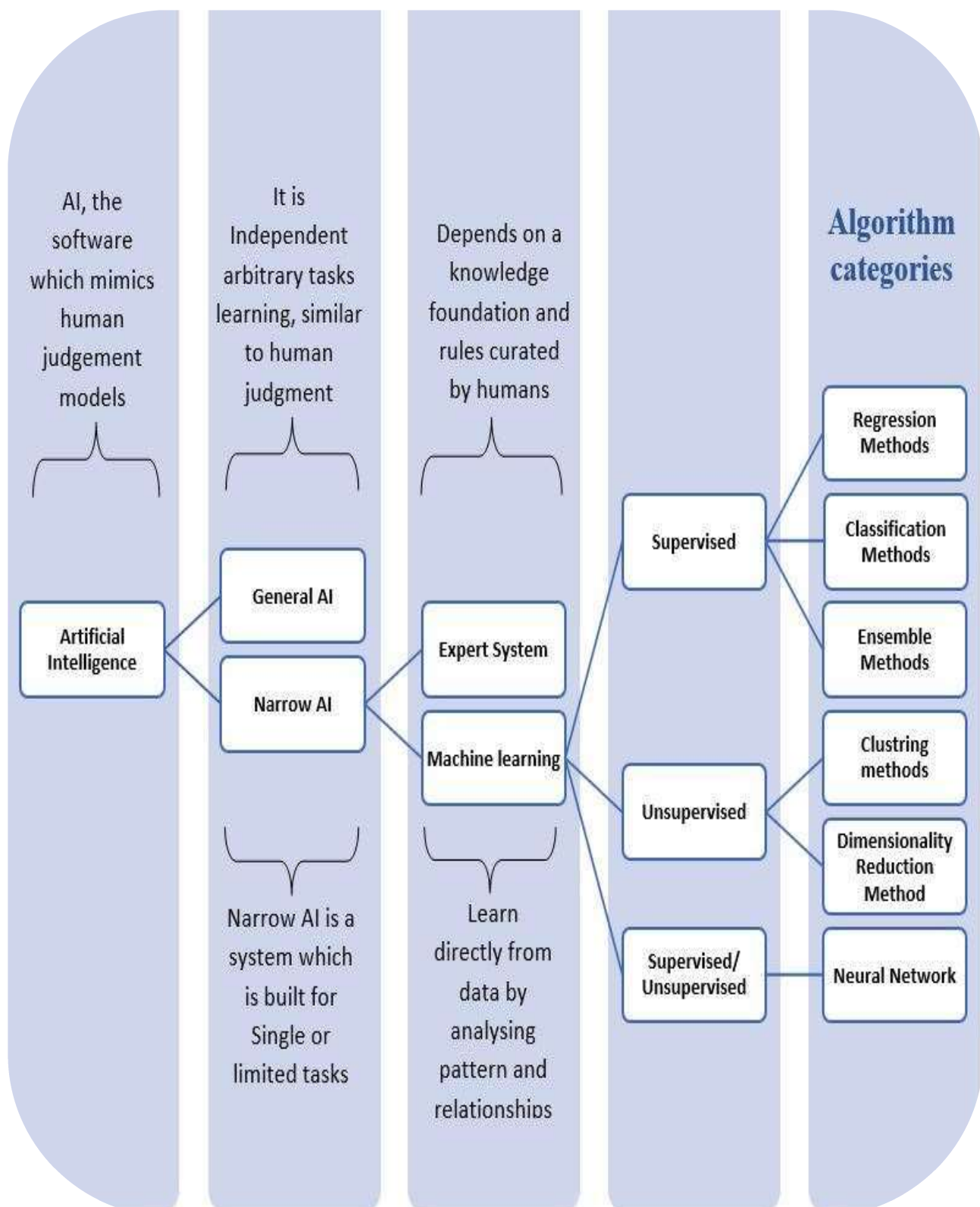


Figure 1 Overview of different type of AI approaches and various algorithm categories.

3 Deep Neural Networks

Deep neural networks are based on ANN (Artificial Neural Network) and is broad class which comes under machine learning. DL approach is mostly used and being adopted by digital pathology. DL easily learn representations directly from primary data and not necessarily depend on engineered features. They are widely accepted because of high accuracy and easier application and hence preferred over hand-crafted feature engineering.

There are various Deep Learning algorithm have been developed like *Convolutional Neural Network (CNN)*; *Fully Convolutional Network (FCN)*; *Generative Adversarial Network (GAN)*; *Recurrent Neural Network (RNN)*.

3.1.1 Convolutional Neural Network (CNN)

Neural networks of these categories are used for Visual imaging analysis and because of this important feature they are the most enormously used DL algorithm for image analysis applications [5]. Based on their shared weight architecture of filters they are also called Space invariant artificial Neural Network (SIANN). The building block of CNN is convolutional sheets where the network learns and in between the input and output layer, there are filters that extract features of an image by analysing the grid-like topology. The image comprises various pixels patches which are arranged in a grid-like manner. CNN DL approach is used for Various tasks e.g., identification and quantification of cells (neutrophil, Lymphocytes, and blast cell). Tschandl et al. [7] organized a study that aims to compare the diagnostic accuracy between machine learning algorithms and expertise human readers. It is the open web-based diagnostic study in which 30 dermatoscopic image batches were selected randomly out of 1511 test set images has been asked to

diagnose and then this diagnosis has been compared with 139 algorithms which had been created by 77 ML laboratories. The interpretation of this study goes in the favor of State-of-the-art machine-learning classifiers as they perform better than human readers. The DL algorithm has the more correct diagnosis and the mean of more correct diagnosis for the DL algorithm is 2.01 (17.91 vs 19.92; $P < 0.0001$) which is obtained by comparing human readers (n=511, human readers) against DL algorithms (n=139, algorithms), but they mentioned a limitation with algorithm performance because the performance of algorithm decreases for out-of-distribution images. Esteva et al. [24] demonstrated skin lesions classification by using a CNN approach. 129,450 clinical images were used by them which contains 2,032 different diseases that trains CNN model. The model is designed in such a way that it can differentiate malignant melanoma from benign nevi and keratinocyte carcinoma from benign seborrheic keratosis. The CNN model performance was compared with 21 board-certified dermatologists. N. Coudray et.al.[16] developed CNN model named “inception v3” which classify Adenocarcinoma, Squamous cell carcinoma, and normal lung tissue accurately on WSIs which are provided by Cancer Genome Atlas. The performance of the model is founded with an average of 0.97 AUC.

3.1.2 Fully Convolutional Neural Network (FCN)

FCN model potentially detects scanted quality pathological images by learning representations from every single pixel. It contains a convolutional layer hierarchy and in place of the last fully connected layer, it has a broad receptive region. Rodner et al.[27] developed FCN based image analysis algorithm. This algorithm can differentiate regions of malignancy from non-malignant regions of epithelium. This differentiation is based on non-linearity found in microscopic images of cancer of the head and neck. Head and neck cancerous sections were co-registered with multimodal images using microscopy. The

WSIs are analysed and segmented into 4 categories malignant epithelium, Non-malignant epithelium, background, and other tissues. They analysed 114 WSIs which is obtained from 12 patients. The average recognition rate obtained is 88.9% and the overall recognition rate obtained is 86.7 of all of the four classes.

3.1.3 Generative adversarial network (GAN)

It is generative modeling that learns from regularities and patterns within the input data.[23] In GAN two neural networks are in a competition with one another. simultaneously in a zero-sum game. One network is the “generator” and another one is the “discriminator”, Generator one generates synthetic data while discriminator one compares the generated and original data for agreement. GAN is used by Gadermayr et al. who performs segmentation of glomeruli and segmented glomeruli out of WSIs from specimens of renal pathology which is isolated from resected mouse kidneys using GAN network. They selected mouse kidney because they are highly similar to human kidney. A. Kapil et al. [28] works on a deep learning model (GAN) for automated estimation of PDL1 and make the first automated scoring GAN model for PDL1 expression.

3.1.4 Recurrent Neural Network (RNN)

RNN is capable of storing input data on different time points for their sequential processing and learns from millions of discrete data [13]. RNN exhibits dynamic behavior as it stores input at a different time interval which is not seen in CNN and FCNs neural networks. CNN learns tasks by analyzing propagated errors and Long Short-Term Memory (LSTM) [14] plays a crucial role in this concern. LSTM is an RNN type that has recurrent gates or forget gates and these gates help CNN in task learning. A network which is been built by the combination of CNN and LSTM by Bychkov et al. [15]. This model analyses H&E-stained tissue of colorectal cancer Thrombotic microangiopathy

(TMA) slide and predict disease re-occurrence. TMA images deconstructed small patches and then data is transferred into CNN. The small patches are easily understood and learned by the model. This combination (CNN+LSTM) gives the predictive performance of $0.69 \text{ AUC} > 0.57 \text{ AUC}$ of histological grade alone and the average score of visual risk (agreed by 3 pathologists) is 0.58 AUC [15].

TABLE III VARIOUS ADVANTAGES AND DISADVANTAGES OF DNN MODELS

DNN MODEL	ADVANTAGE	DISADVANTAGE
CONVOLUTIONAL NEURAL NETWORK (CNN)	High Accuracy; Automated extraction of features; fast to train	Large dataset requires; Explainability is low (CNN Black box)
(FULLY CONVOLUTIONAL NETWORK)	Splits image of any size; shape, size and position of target is analyzed using pixel-level detection	Labelling cost is high
GENERATIVE ADVERSARIAL NETWORK	Labeling isn't required; It learns and synthesizes realistic data	Training is complex and unstable.
RECURRENT NEURAL NETWORK	learns sequential Data (different time intervals)	Computational cost is high

4 AI Publications Boom In Modern Era

After technological advancements and the introduction of Digital Pathology, the market has grown remarkably with an exponential curve in the graph. The AI publication number drastically increased by more than 6 times within the last 20 years of the span. The Healthcare sector also came up with innovative ideas and starts utilizing or synergizing the computer science tool in their domain. Hence in result, the publications in biomedical literature have increased more than 8 times since 2000 and the Global AI market has estimated to be increased to \$300 billion from \$100 billion by 2025 and growth rate will be exponential with growth rate of around 40-50 % and Global AI in healthcare market is estimated to increase from \$856 million to \$20 Billion by the year 2025 [4]. Y Guo et al.[29] performed a bibliographic analysis of AI publications in healthcare. For this, they developed a strategy for searching published papers in the AI-related health care category. For this, the research papers including citations related to AI in healthcare data was retrieved from Web of Science (Wos). the publication records up to December 2019 were noted and a total of 5235 papers founded. From 5235 papers, 32 were excluded which are duplicates and currently proceeding, and 3730 were excluded because of not fulfilling screening criteria, so at last 1473 papers contributed to bibliometric analysis and steeply growth was seen from 2015-2019 with 987/1456 publications (67.78%). The timeline of the publications was given below (Table I) along with publications growth graph (Fig.2.)

TABLE IV TIMELINE SHOWING THE NUMBER OF AI PUBLICATIONS AND ITS CONTRIBUTION PERCENTAGE IN THE AI-RELATED HEALTHCARE PUBLICATIONS.

YEAR OF PUBLICATION	NUMBER OF PUBLICATION	CONTRIBUTION PERCENTAGE %
1995-1999	38	2.60 %
2000-2004	97	6.66 %
2005-2009	98	6.73 %
2010-2014	236	16.20 %
2015-2019	987	67.78%
TOTAL PUBLICATIONS	1456	100%

^a. Data source: - Data aggregated from study of Y Guo et al. [29] <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7424481/>.

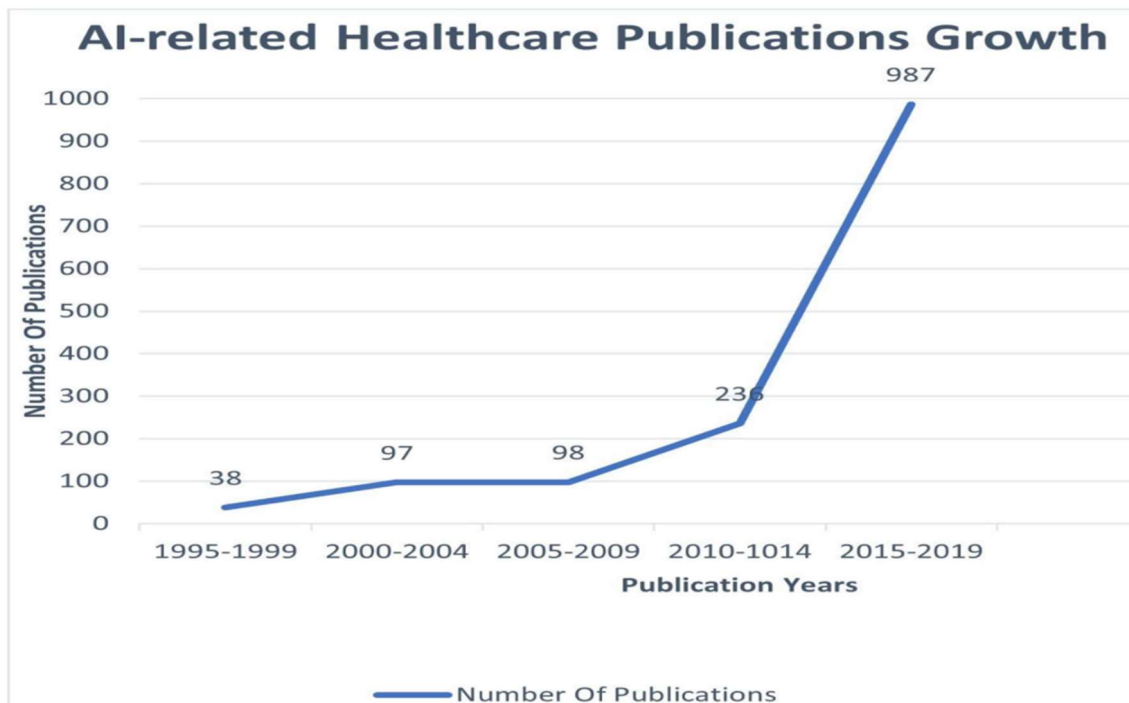


Figure 2 AI publication curve

5 Application of AI synergy in Healthcare

5.1 Application of AI in Pathology: -

Pathology is the most important branch of Medical Science which fundamentally concerns the origin, cause, and nature of the disease. This domain contributes the most important part of the diagnostic infrastructure. The pathological examination includes examining, Tissues, Organs, Body fluids. In this examination, the tissue is fixed on a glass slide by certain procedures and then observed under a microscope for disease. Now with technological advancement, the traditional pathological operation is getting evolved and emerged as Digital pathology. Digital pathology comprises the digitalization of Histopathological slides using a Whole-slide scanner which is a microscope under robotic and computer control and analysing these digitalized images or whole-slide image using a Computational approach. This digitalization creates a high-resolution, enhanced pixel image of pathological specimen that ultimately generates a tremendous amount of data up to gigabytes per biopsy.

Because of algorithmic advancement and more convenient computing power, in the past few decades, Artificial Intelligence and machine learning have spread their roots into healthcare and clinical applications. And artificial intelligence and digital pathology synergy is one of the examples of these two fields intersection. This synergy will have majestic results in Diagnostic, prognostic, and predictive analysis of Whole-slide images.

5.1.1 How Does It work:

Whole slide scanners form a digital image of histopathological slides and then AI technology is applied to these images for image processing and classification task in which the AI technology performs low-level tasks and high-level tasks respectively, In low-level tasks, object recognition problem is mainly focused like detection and segmentation and High-level tasks mainly focused on prediction of disease diagnosis and prognosis of disease by just analysing the digitalized pattern of the image. AI algorithms were designed in such a way that they can access and analyse the slide for certain tumors type and hence many AI algorithms need to design in order to get an accurate pathological report, in which tumors in tissue need to be diagnosed, graded, subtyped and accessing of high risk features should be done by another individual algorithm.

Various AI techniques such as machine learning, deep learning are used to convert WSI images into mineable data and these computation techniques come under Machine Vision. Machine vision is the technology that performs to extract information directly from an image using imaging-based analysis or inspection. By this approach interpretation of digital biopsies, samples could be done as quantitative datasets. Primary data performs a major role in AI performance. Quality of primary data decides AI performance. In order to achieve maximum prediction performance, primary data should have comprehensiveness, accuracy, and cleanliness.

5.2 In Bioinformatics:

In the discipline data management systems such as MapReduce and Hadoop are the most extensively utilised to provide better data warehouse repositories, computational infrastructure, and data mining tools to evaluate biological data in a manageable time

period. Clinical genomic analysis today employs both open source and commercial methods to decipher the hidden meanings in genetic data.

However, using these tools to create value from data is difficult due to a lack of expertise and experience. Organizations must choose the proper tool from among the numerous tools available on the market to solve their business difficulties.

5.3 In Medical Imaging Technology.

In the field of medical imaging, reports made utilising deep learning techniques assist clinicians in better analysing life-threatening and malignant disorders than in the past.

5.4 In pharmaceuticals

In pharmaceuticals AI could be used in Manufacturing process and improvement tasks. Drug discovery and design is one of the major advantages of this technology. Target prediction and bioactivity analysis property of AI tool has changes the traditional drug discovery workflow.

Processing of biomedical and clinical data is now work of ease after implication of AI in pharmaceuticals.

6 Drug Discovery and Regression Models Accuracy Analysis Using Machine Learning Algorithms

Machine learning is one of the most important and rapidly growing technology in computer-aided drug discovery. Machine learning methods are far more successful than physical models, allowing them to scale to enormous datasets without requiring a lot of processing power. One of the most prominent applications of machine learning in drug

development is to help researchers identify and exploit chemical-biological activity correlations, often known as SAR.

The biological activities of compounds that are of interest in drug development have been discovered using machine learning and statistical techniques. Machine learning techniques are used to construct various quantitative structure activity relationship (QSAR) models for understanding bioactivity of substances. In this study we perform various statistical analysis and validate QSAR model for categorize various regression algorithm on the basis of their performance. The performance parameters will be correlation coefficient, R squared , Adjusted R square and Root Mean Square Deviation (RMSD)

6.1 Literature Review

Rheumatoid arthritis is a autoimmune disease of joints that is caused by chronic inflammation in joints. This inflammation cause joint damage and lead to disability of that for lifelong. 0.5-1 % of adults got affected annually with RA in industrialized countries. Risk factor of RA could be environmental or genetics. Genetics contributes of about 50% for a risk in order to a person develop RA. RA typically affects women's and elderly peoples. Due to medical advance RA clinical status in recent years has been improved. Advances in diagnosis and treatment activity of the disease could be reduce. Disease-Modifying Anti Rheumatic Drugs (DMARDs) has shown promising result in RA. DMARDs are targeted synthetic drugs. They suppress inflammation or overactive immune system and provide relief to RA patients. They may take several weeks or months for being effective in RA. RA has no complete cure and patients have to rely on DMARDs. T cell, B cells and pro-inflammatory cytokines plays major role in RA

pathophysiology Various pro-inflammatory cytokines such as Tumor Necrosis Factor (TNF- α), IL-6, IL7, IL1 β , IL17, IL18, IL23, Granulocyte Macrophage Colony-Stimulating Factors (GM-CSF) AND Gamma interferon (IFN)- γ are involved in causing inflammation in Rheumatoid arthritis. Tumor Necrosis Factor (TNF- α), Interleukin 6 (IL-6) and Interleukin 1 (IL-1) plays major role in causing synovial inflammation. Because of direct involvement of cytokines in RA, they have been explore as potential targets to dampen inflammation process. Use of TNF inhibitors is one of the important and approved target specific treatment of Rheumatoid arthritis. The inflammation signalling pathway is ceased by TNF inhibitors and provides potential treatment for the disease.

In this thesis Machine learning based regression models have been used to understand the interaction of biological activity of compounds to TNF alpha inhibitors.

I have compared various regression model for the accuracy.

6.2 Methodology

In order to understand correlation between biological activity of compound and the target. It is very important to have a clean and quality dataset. There are various databases available online from where data could be gathered. Preparation of the dataset for the mentioned thesis is as follow:

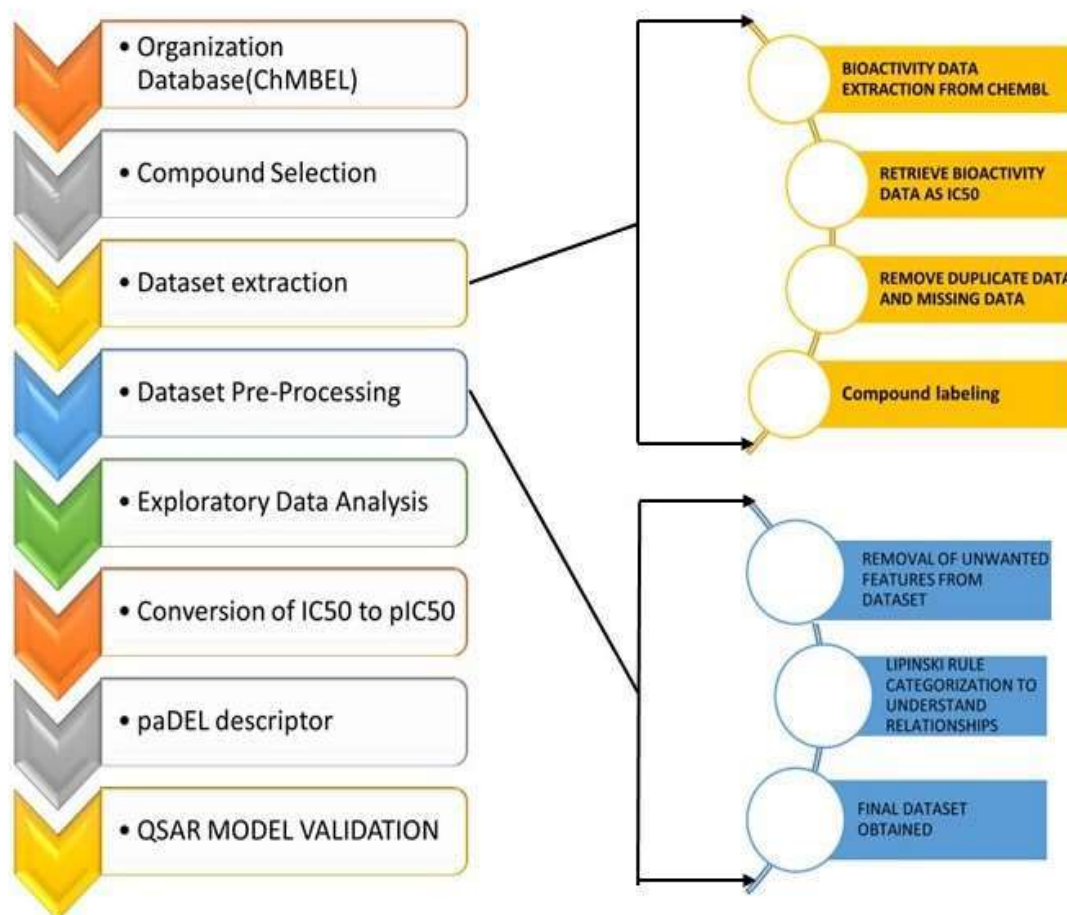


Figure 3 Methodology of thesis

6.2.1 Biological Data Extraction using ChEMBL Database:

TNF-alpha is searched for in the ChEMBL database to locate all possible target chemicals.

ChEMBL (<https://www.ebi.ac.uk/chembl>)[34] is a manually curated library of bioactive compounds with drug-like properties. It combines chemical, bioactivity, and genetic data to help translate genetic data into effective new drugs. There are 1231 targets in the collection.

TNF-alpha bioactivity data for Homo sapiens is acquired using the unique ChEMBL ID

"ChEMBL1825." Bioactivity data is a type of scientific data that must be searchable, accessible,

interoperable, and reusable in order to be useful. One of the most essential properties of chemical compounds is pharmacological/biological activity, which shows how the chemicals might be utilised in medicinal applications.

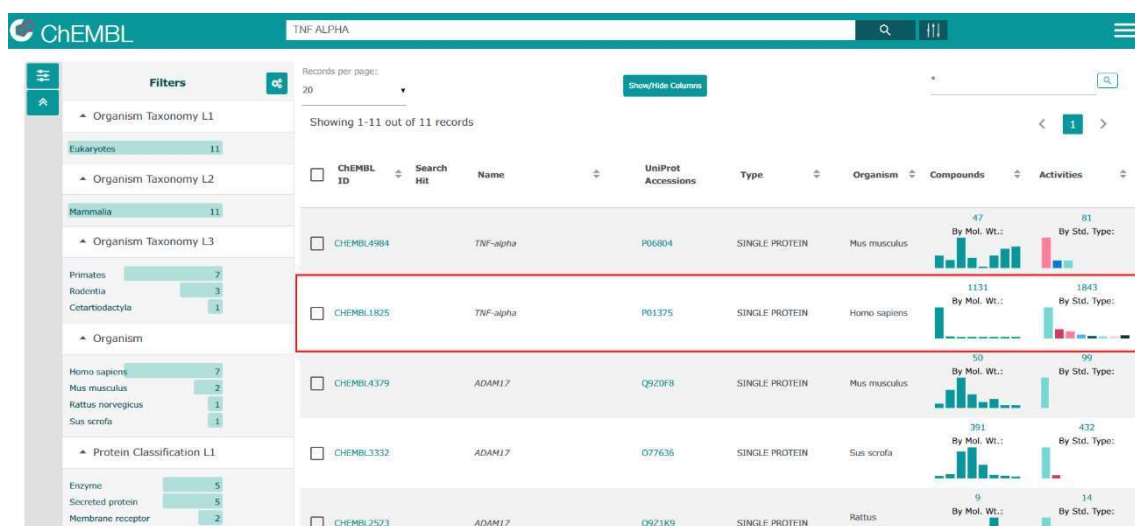


Figure 4 Retrieving bioactivity data from ChEMBL

6.2.2 Retrieve Bioactivity data reported as IC50 values in nM (nanomolar) unit:

There are various bioactivity measurement units available at the ChEMBL database but we retrieve IC50 bioactivity measurement unit is made up major subset of the compounds. The IC50 of an inhibitor is the concentration at which the response (or binding) of the inhibitor is halves. The standard value represents the drug's potency; the lower the value, the more potent the substance. For an idealistic state, the standard value should be as low as possible, i.e. the inhibitory concentration at 50% should be low. This means that a lower medication concentration would be required to block the target protein 50% of the time. As a result, only the results presented as IC50 values in nM (nanomolar) units were retrieved.

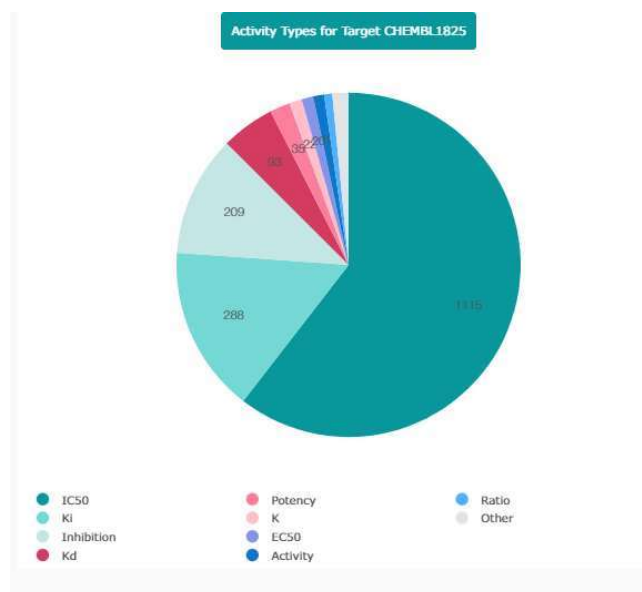


Figure 5 Pie chart showing major subset of IC50

6.2.3 Remove Duplicate Data and Missing data:

The standard value and canonical smiles columns of the dataset is examined for any missing values. Before the molecules are labelled, the missing values are removed. This was done to make the data more uniform so that it could be analysed further. It was possible to obtain a high-quality dataset of 959 chemicals. For the classic grins notation, numerous compounds had the same values. In order to have only unique values in the final dataset, these redundant values were deleted.

6.2.4 Compound Labelling :

Compounds have their IC50 value in dataset and this value is categorized as Active, Intermediate, Inactive. This classification of data is done on the basis of IC50 values. The compounds having IC50 values $\geq 10,000\text{nm}$ are classified as inactive class, The compounds having values $\leq 1000\text{nm}$ are classified as active class, and The compounds having values in between to that of 1000 and 10000 were considered as intermediate class.

6.2.5 Pre-Processing of Datasets

Data pre-processing is a crucial step for determining accurate interpretations. For this I go to following procedure of data pre-processing:

6.2.5.1 Removal of unwanted features from datasets:

From datasets certain bioactivity measurement units were removed and keep those columns containing features like Molecule ChEMBL ID, Canonical Smiles, Bioactivity Class and Standard Value.



	A	B	C	D
1	Molecule ChEMBL ID	Canonical Smiles	Bioactivity class	Standard Value
2	CHEMBL3640303	<chem>Cn1cc(-c2ccc3c(c2)C(=O)N(C[C@@]2(C#Cc4ccccc4)NC(=O)NC2=O)C3)cn1</chem>	Inactive	10000001
3	CHEMBL3640304	<chem>CSc1ccc2c(c1)C(=O)N(C[C@@]1(C#Cc3ccccc3)NC(=O)NC1=O)C2</chem>	Active	651.1
4	CHEMBL3640323	<chem>COc1ccc2c(c1)C(=O)N(C[C@@]1(C#Cc3ccc(C(CN)N4CCC(N)C4)cc3)NC(=O)NC1=O)C2</chem>	Active	233
5	CHEMBL3640325	<chem>COc1ccc2c(c1)C(=O)N(C[C@@]1(C#Cc3cnccc(-c4ccc(O)c(-c5cn[nH]c5)n4)c3)NC(=O)NC1=O)C2</chem>	Active	374
6	CHEMBL3640336	<chem>COc1ccc2c(c1)C(=O)N(C[C@@]1(C#Cc3ccc(-c4nc(-c5ccccc5)ccc4O)cc3)NC(=O)NC1=O)C2</chem>	Active	65
7	CHEMBL3640342	<chem>CCN1CCN(Cc2ccc(C#C[C@@]3(CN4Cc5ccc(OC)cc5C4=O)NC(=O)NC3=O)cn2)CC1</chem>	Active	782.3
8	CHEMBL3640361	<chem>COc1ccc2c(c1F)C(=O)N(C[C@@]1(C#Cc3ccc(-c4cccc(N5CCOCC5)n4)c(F)c3)NC(=O)NC1=O)C2</chem>	Intermediate	2557
9	CHEMBL3640365	<chem>COc1ccc2c(c1)C(=O)N(C[C@@]1(C#Cc3ccc(-c4ccc(-c5cccc(N6CCSCC6)n5)cc4)cc3)NC(=O)NC1=O)C2</chem>	Inactive	1000001
10	CHEMBL3640369	<chem>COc1ccc2c(c1F)C(=O)N(C[C@@]1(C#Cc3ccc(-c4nc(-c5cnn(C)c5)ccc4OC(=O)C(C)(C)c(F)c3)NC(=O)NC1=O)C2</chem>	Active	68.45
11	CHEMBL3640372	<chem>COc1ccc2c(c1)C(=O)N(C[C@@]1(C#Cc3ccc(-c4cncnc4C)nc3)NC(=O)NC1=O)C2</chem>	Intermediate	2392
12	CHEMBL3640376	<chem>COc1ccc2c(c1)C(=O)N(C[C@@]1(C#Cc3ccc(-n4cnc(CCN(C)=O)c4)nc3)NC(=O)NC1=O)C2</chem>	Active	524
13	CHEMBL3640379	<chem>COc1ccc2c(c1)C(=O)N(C[C@@]1(C#Cc3ccc(-n4ccnc4)cc3)NC(=O)NC1=O)C2</chem>	Active	786
14	CHEMBL3640381	<chem>COc1ccc2c(c1)C(=O)N(C[C@@]1(C#Cc3ccc(-c4cccc(C)c4O)nc3)NC(=O)NC1=O)C2</chem>	Intermediate	3505
15	CHEMBL3640382	<chem>COc1ccc2c(c1)C(=O)N(C[C@@]1(C#Cc3ccc(-c4ccc(C)[nH]c4=O)nc3)NC(=O)NC1=O)C2</chem>	Active	766
16	CHEMBL3640389	<chem>COc1ccc2c(c1F)C(=O)N(C[C@@]1(C#Cc3ccc(C(N=O)N4CCNC(C@@H)4C)c(F)c3)NC(=O)NC1=O)C2</chem>	Active	42
17	CHEMBL3640391	<chem>CCN1CCN(C(N=O)c2ccc(C#C[C@@]3(CN4Cc5ccc(OC)c(F)c5C4=O)NC(=O)NC3=O)cc2F)[C@@H](C)C1</chem>	Active	69
18	CHEMBL3640392	<chem>COc1ccc2c(c1F)C(=O)N(C[C@@]1(C#Cc3cc(F)enc3C)NC(=O)NC1=O)C2</chem>	Active	935
19	CHEMBL3640402	<chem>COc1ccc2c(c1)C(=O)N(C[C@@]1(C#Cc3ccc(N4C(=O)CCC4C(=O)NC4CC4)cc3)NC(=O)NC1=O)C2</chem>	Active	127
20	CHEMBL3640403	<chem>COc1ccc2c(c1F)C(=O)N(C[C@@]1(C#Cc3ccc(N4C(=O)CCC4C(=O)NC4CC4)cc3)NC(=O)NC1=O)C2</chem>	Active	278
21	CHEMBL3640415	<chem>COc1ccc2c(c1)C(=O)N(C[C@@]1(C#Cc3ccc(-n4c(-c5ccccc5)n[nH]c4=O)cc3)NC(=O)NC1=O)C2</chem>	Active	258

Figure 6 Dataset after removal of unwanted features

6.2.5.2 Lipinski rule categorization to understand relationships:

Christopher Lipinski, developed a set of standards for determining whether or not a chemical is drug-like. To identify drug similarity, the Absorption, Distribution, Metabolism, and Excretion (ADME) profile, also known as the pharmacokinetic profile, is used. Lipinski developed the Rule-of-Five, often known as Lipinski's Rule, after evaluating all orally active FDA-approved drugs. Any medicine that meets two or more of the criteria is predicted to have low permeability and absorption, according to the guideline. The rule may be able to classify and

eliminate compounds with a low drug-likeness prediction. In lipinski's descriptors the data filtration occurs on the basis of certain parameters e.g.

- i) Molecular weight greater than 500 Da, filtered.
- ii) Hydrogen bond donors more than 5, filtered
- iii) Hydrogen acceptors more than 10, filtered
- iv) Log P values above 5, filtered

These above mentioned are the threshold values for lipinski's descriptors. These values are used studying and observing statistical distribution. After calculating descriptors the data is merged into the datasets.

TABLE III DATASET VALUES AFTER FILTERING DATA ON THE BASIS OF LIPINKSI'S RULE

Molecular Weight	LOG p	Hydroge acceptor	Hydrogen donor
426.44	1.07	9	2
392.44	1.39	7	2
425.42	1.04	9	3
472.5	1.06	9	3
313.36	2.32	6	0
435.91	4.87	6	0
445.44	0.46	10	3
323.4	3.15	5	0
325.37	2.37	6	0
478.94	3.55	8	0
462.51	0.65	10	3
459.51	1.67	9	2
472.42	1.84	8	4
456.46	1.08	10	2
487.56	1.56	9	2
468.47	2.05	9	3
456.46	1.25	10	5
484.52	1.69	10	2
445.56	3.27	6	0
397.62	3.75	7	2

6.2.5.3 Exploratory Data Analysis:

The data subset comprising the descriptor values is then subjected to graphical and statistical analysis in order to gain a better understanding of the data points via visualisations. EDA is a vital step in undertaking early investigations on data for revealing patterns, finding anomalies, testing hypotheses, and probing for assumptions by using summary statistics and graphical representations.

6.2.5.4 Final dataset obtained after pre processing

	A	B	C	D	E	F	G	H
	Molecule ChEMBL ID	Molecular Weight	Canonical Smiles	Bioactivity class	Standard Value	Log P	Hydroge acceptor	Hydrogen donor
1	CHEMBL3640303	426.44	Cn1cc(-c2ccc3c(c2)C(=O)N[C@H]2[C@@H](C#Cc4ccnc4)NC(=O)NC2=O)C3)cn1	Inactive	10000001	1.07	9	2
2	CHEMBL3640304	392.44	CSc1ccc2c(c1)C(=O)N[C@H]1[C@@H](C#Cc3ccnc3)NC(=O)NC1=O)C2	Active	651.1	1.39	7	2
4	CHEMBL3640323	502.58	COc1ccc2c(c1)C(=O)N[C@H]1[C@@H](C#Cc3ccc(CCN)N4CCCC(NC4)cc3)NC(=O)NC1=O)C2	Active	233	0.31	10	6
5	CHEMBL3640325	535.52	COc1ccc2c(c1)C(=O)N[C@H]1[C@@H](C#Cc3ccc(-c4ccc(O)c(-c5cn[nH]c5)n4)3)NC(=O)NC1=O)C2	Active	374	1.83	12	4
6	CHEMBL3640336	545.56	COc1ccc2c(c1)C(=O)N[C@H]1[C@@H](C#Cc3ccc(-c4nc(-c5ccnc5)ccc4)cc3)NC(=O)NC1=O)C2	Active	65	3.11	10	3
7	CHEMBL3640342	502.58	CCN1CCN(CC2ccc(C#C[C@H]3(CN4Cc5ccc(OC)cc5C4=O)NC(=O)NC3=O)cn2)CC1	Active	782.3	0.81	10	2
8	CHEMBL3640361	573.56	COc1ccc2c(c1F)C(=O)N[C@H]1[C@@H](C#Cc3ccc(-c4ccc(N5CCOCC5)n4)c(F)3)NC(=O)NC1=O)C2	Intermediate	2557	2.46	10	2
9	CHEMBL3640365	629.74	COc1ccc2c(c1)C(=O)N[C@H]1[C@@H](C#Cc3ccc(-c4ccc(-c5cccc(N6CCSCC6)n5)cc4)cc3)NC(=O)NC1=O)C2	Inactive	10000001	4.56	9	2
10	CHEMBL3640369	668.66	COc1ccc2c(c1F)C(=O)N[C@H]1[C@@H](C#Cc3ccc(-c4nc(-c5cnn(C)k5)ccc4OC(=O)C(C)C(F)3)NC(=O)N	Active	68.45	3.97	9	2
11	CHEMBL3640372	467.49	COc1ccc2c(c1)C(=O)N[C@H]1[C@@H](C#Cc3ccc(-c4cccc4)nc3)NC(=O)NC1=O)C2	Intermediate	2392	2.05	9	2
12	CHEMBL3640376	527.54	COc1ccc2c(c1)C(=O)N[C@H]1[C@@H](C#Cc3ccc(-n4nc(CCN(C)C=O)k4)nc3)NC(=O)NC1=O)C2	Active	524	0.54	12	3
13	CHEMBL3640379	441.45	COc1ccc2c(c1)C(=O)N[C@H]1[C@@H](C#Cc3ccc(-n4cnc4)cc3)NC(=O)NC1=O)C2	Active	786	1.47	9	2
14	CHEMBL3640381	482.5	COc1ccc2c(c1)C(=O)N[C@H]1[C@@H](C#Cc3ccc(-c4cccc(C)k4)nc3)NC(=O)NC1=O)C2	Intermediate	3505	2.36	9	3
15	CHEMBL3640382	483.48	COc1ccc2c(c1)C(=O)N[C@H]1[C@@H](C#Cc3ccc(-c4ccc(C)[nH]k4=O)nc3)NC(=O)NC1=O)C2	Active	766	1.34	10	3
16	CHEMBL3640389	552.54	COc1ccc2c(c1F)C(=O)N[C@H]1[C@@H](C#Cc3ccc(C(N=O)N4CCN[C@H]4C)k(F)3)NC(=O)NC1=O)C2	Active	42	1.62	11	3
17	CHEMBL3640391	580.59	CCN1CCN(C(N=O)k2ccc(C#C[C@H]3(CN4Cc5ccc(OC)C(F)5C4=O)NC(=O)NC3=O)cc2F)[C@H](C)C1	Active	69	2.35	11	2
18	CHEMBL3640392	446.8	COc1ccc2c(c1F)C(=O)N[C@H]1[C@@H](C#Cc3ccc(F)cn3)C1)NC(=O)NC1=O)C2	Active	935	1.61	8	2
19	CHEMBL3640402	541.56	COc1ccc2c(c1)C(=O)N[C@H]1[C@@H](C#Cc3ccc(N4C(=O)CCCC4(=O)NC4CC4)cc3)NC(=O)NC1=O)C2	Active	127	1.05	11	3
20	CHEMBL3640403	559.55	COc1ccc2c(c1F)C(=O)N[C@H]1[C@@H](C#Cc3ccc(N4C(=O)CCCC4(=O)NC4CC4)cc3)NC(=O)NC1=O)C2	Active	278	1.19	11	3
21	CHEMBL3640415	535.52	COc1ccc2c(c1)C(=O)N[C@H]1[C@@H](C#Cc3ccc(-n4(-c5ccnc5)n[nH]k4=O)3)NC(=O)NC1=O)C2	Active	258	1.22	12	3
22	CHEMBL3642432	529.51	COc1ccc2c(c1)C(=O)N[C@H]1[C@@H](C#Cc3ccc(NC4nccc(=O)[nH]4)cc3OC)NC(=O)NC1=O)C2	Active	632	0.38	13	4
23	CHEMBL3642437	407.39	COc1ccc2c(c1)C(=O)N[C@H]1[C@@H](C#Cc3cnc(N)nc3)NC(=O)NC1=O)C2	Intermediate	1597	-0.76	11	6
24	CHEMBL3642440	485.46	COc1ccc2c(c1)C(=O)N[C@H]1[C@@H](C#Cc3ccc(-c4ccc(C(N)=O)k4)3)NC(=O)NC1=O)C2	Active	776.1	1.03	11	4
25	CHEMBL3642447	432.44	COc1ccc2c(c1)C(=O)N[C@H]1[C@@H](C#Cc3ccc(C(N)=O)ccc3)NC(=O)NC1=O)C2	Active	417	0.69	9	4

Figure 7 Final dataset obtained after pre-processing

6.2.6 IC50 to pIC50 conversion:

The negative logarithmic values of IC50 are known as pIC50 values. The pIC50 values are calculated to make the data more homogeneous for analysis and graph plotting. It will encourage you to think about your potency data in logarithmic scales instead than arithmetic scales. The precision of IC50 determinations is challenging to report.

This conversion initially done by using a software [1] but as it is time taking procedure so I used formula $9 - \log(\text{IC}_{50})$ in excel which gives accurate pIC50 value.

6.2.7 Conversion of structural and chemical properties into numerical values:

The paDEL software [33] is being used in this conversion. This software convert structural and chemical properties into paDEL descriptors which are the numerical values. These numerical data are utilised as input for QSAR model construction and can be used for predicting the bioactivity of new drugs. The descriptor is offered in library format and includes about 43 molecular descriptor techniques as well as seven fingerprint algorithms. It is open source software that is provided for free, making it more accessible for use than other software options.

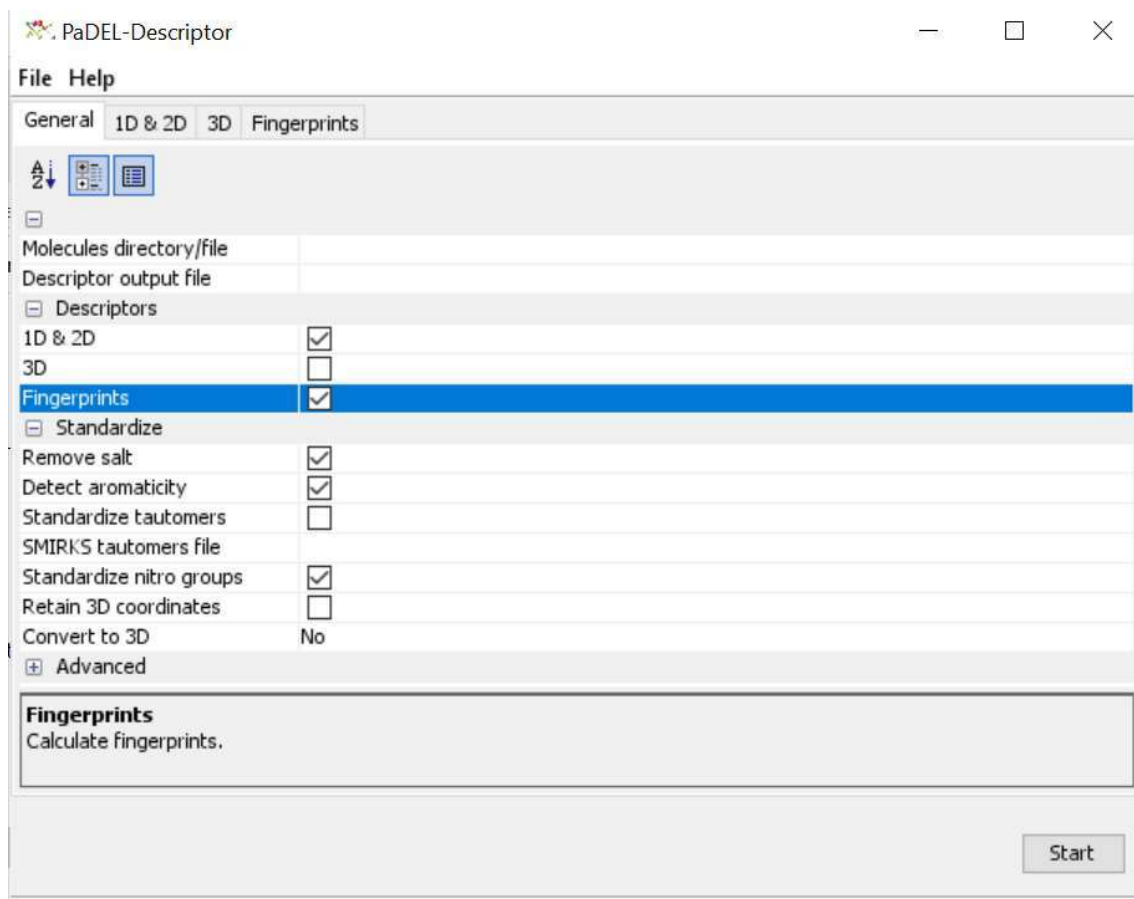


Figure 8 PaDEL- Descriptor software

7 QSAR MODEL VALIDATION USING WEKA

Validating the model that has been constructed is a crucial step in ensuring that the model can make accurate and dependable predictions. Pearson's correlation coefficient (r) and root mean squared error are the most widely utilised parameters to assess the model's performance (RMSE). Pearson's coefficient is a numerical representation of the degree of association between the features under investigation, with values ranging from -1 to +1. A positive correlation is shown by a positive value, and vice versa. The Root Mean Squared Error is a parameter that is used to assess the model's potential for error.[36]

8 Result and Interpretation of EDA and Statistical Analysis

The scatter plot of MW vs LogP (Fig) shows that the two bioactivity classes are found in chemical regions that are essentially identical. The QSAR model takes into account a vast number of substances, each of which can be represented by a set of descriptors. The active and inactive chemicals have a good association according to the chemical space analysis.

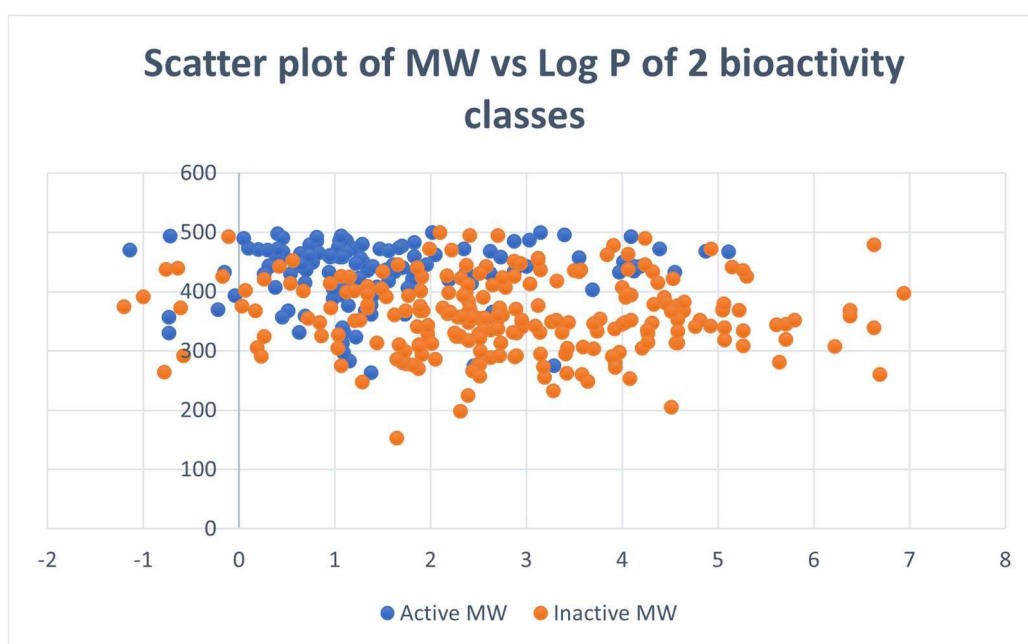


Figure 9 Scatter plot of MW vs Log P of 2 bioactivity classes

The different box plots were made to help comprehend the chemical space of TNFalpha inhibitors and and get outcome about the Specific Absorption Rate (SAR) using Lipinski's Descriptors. This type of chemical space analysis reveals the common properties of the compounds under investigation. These descriptors were subjected to EDA in order to better comprehend the statistical inference of the distribution over the chemical space. The most common measure utilised is molecular weight (MW), because knowing the molecular weight of a molecule is essential to PREDICT its capacity to pass through a lipid bilayer. The active

inhibitors are found to be in the 400-500 Da range that means the the compound with MW in that range will be effective against TNF-alpha.

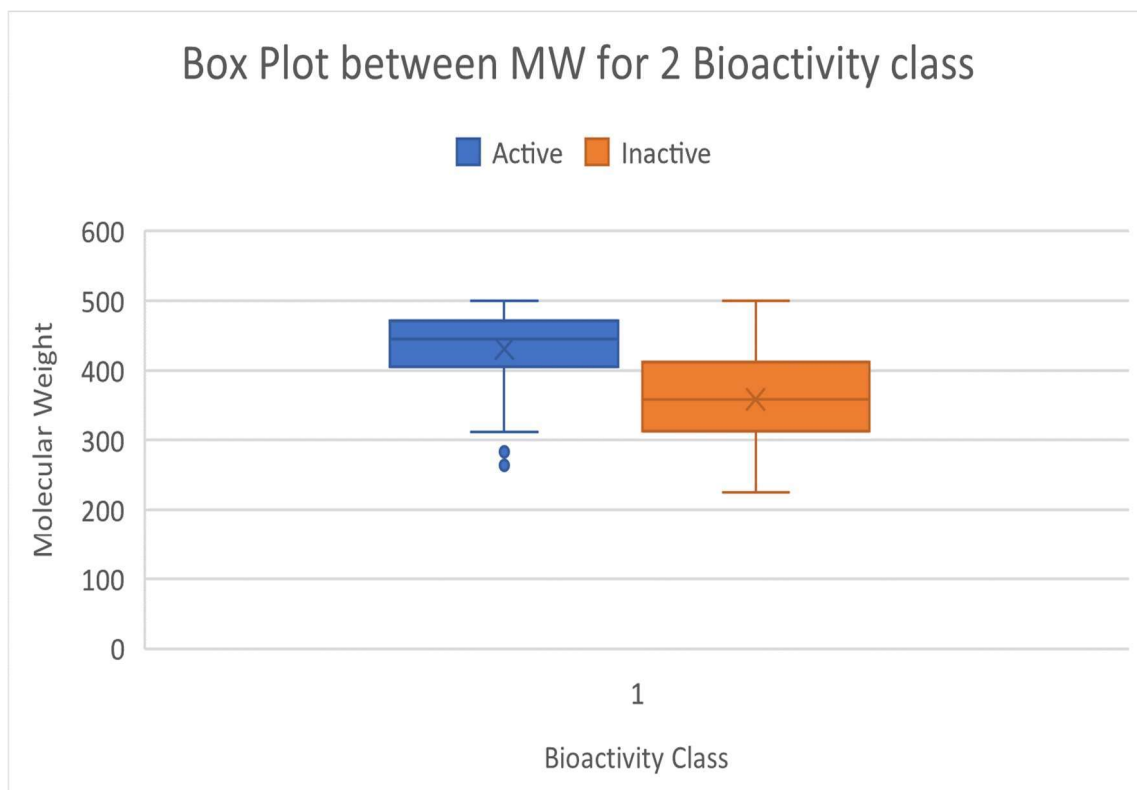


Figure 10 Box plot between MW of 2 bioactivity class

LogP is also a popular method for determining a compound's penetration and permeation capabilities. It provides insight into the lipophilicity of the compounds. The active compounds are distributed over a range of 1 to 3, whereas the inactive ones are distributed over a range of 2 to 4. The hydrogen bonding capacity of the compounds is measured using HB-donors and HB-acceptors.

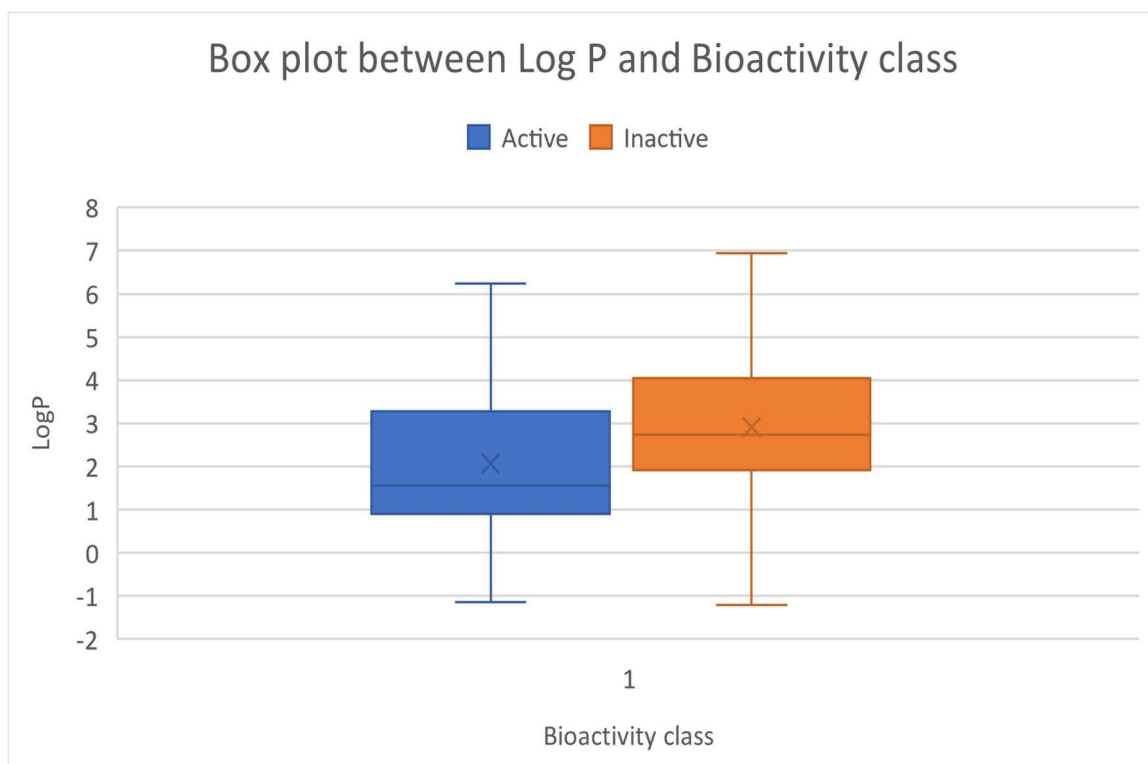


Figure 11 Box plot between Log P of 2 bioactivity class

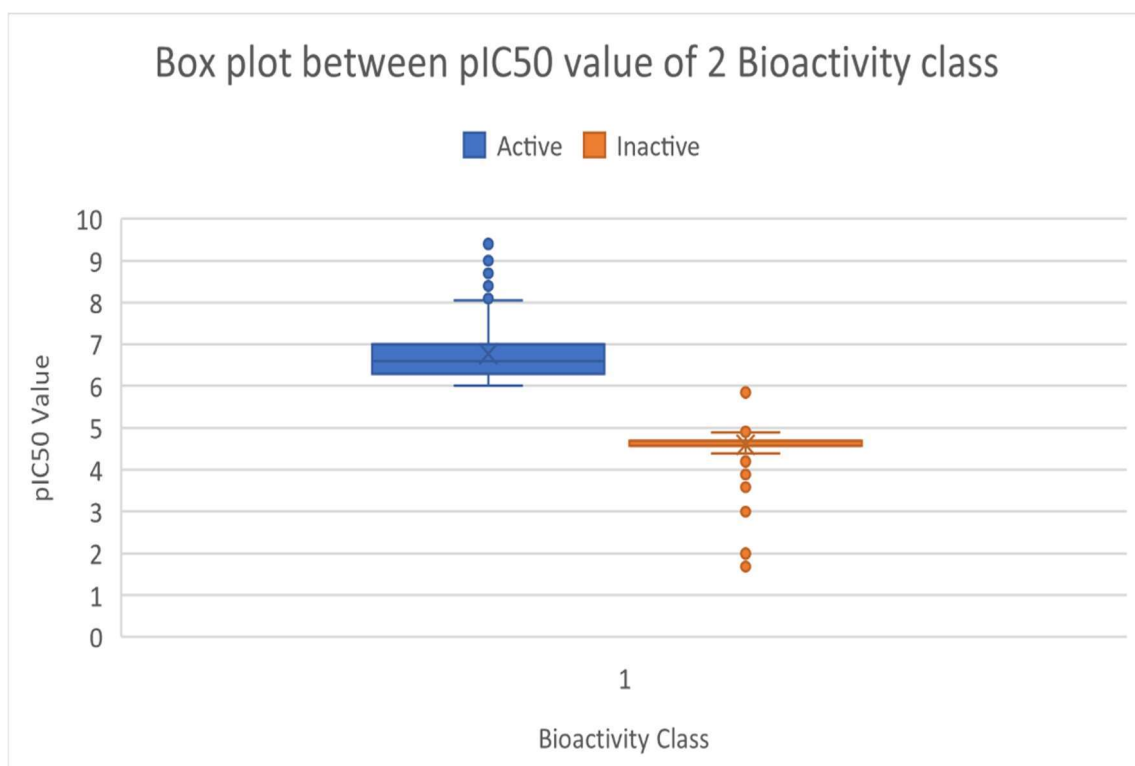


Figure 12 Box plot between pIC50 value of 2 bioactivity class

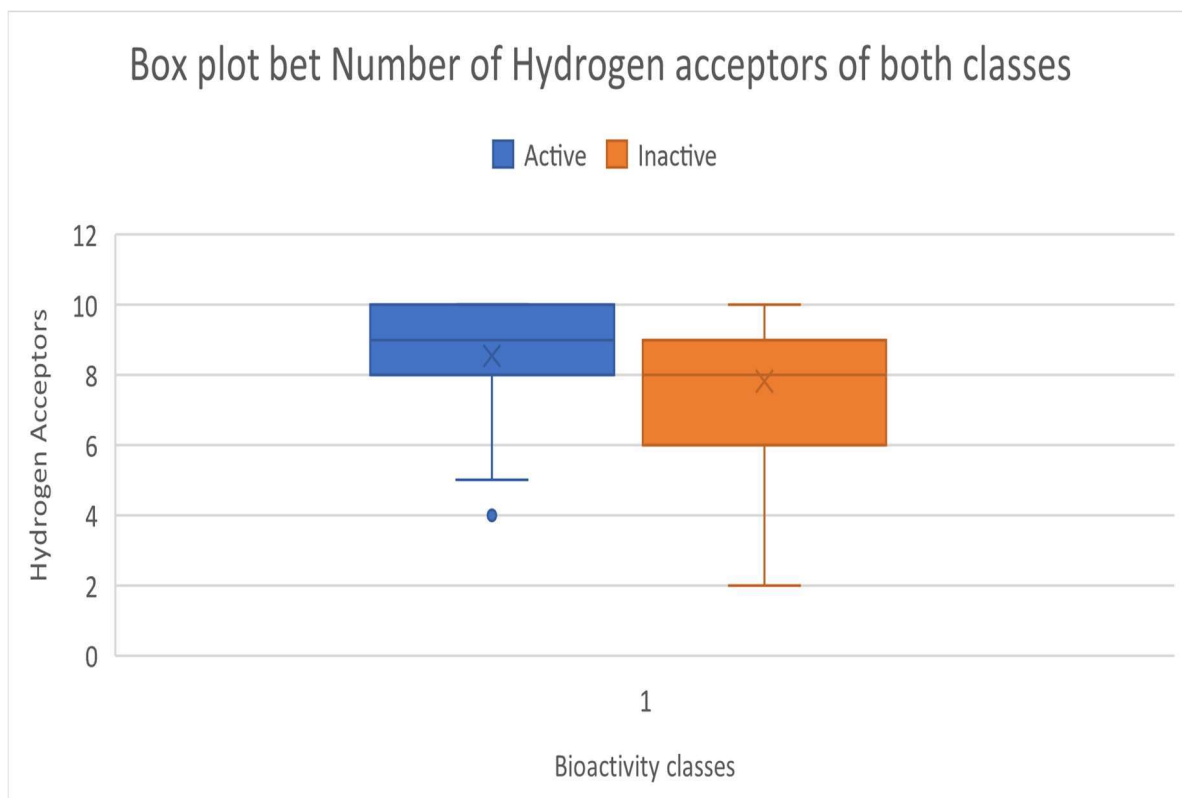


Figure 13 Box pot between Number of hydrogen acceptor of 2 bioactivity class

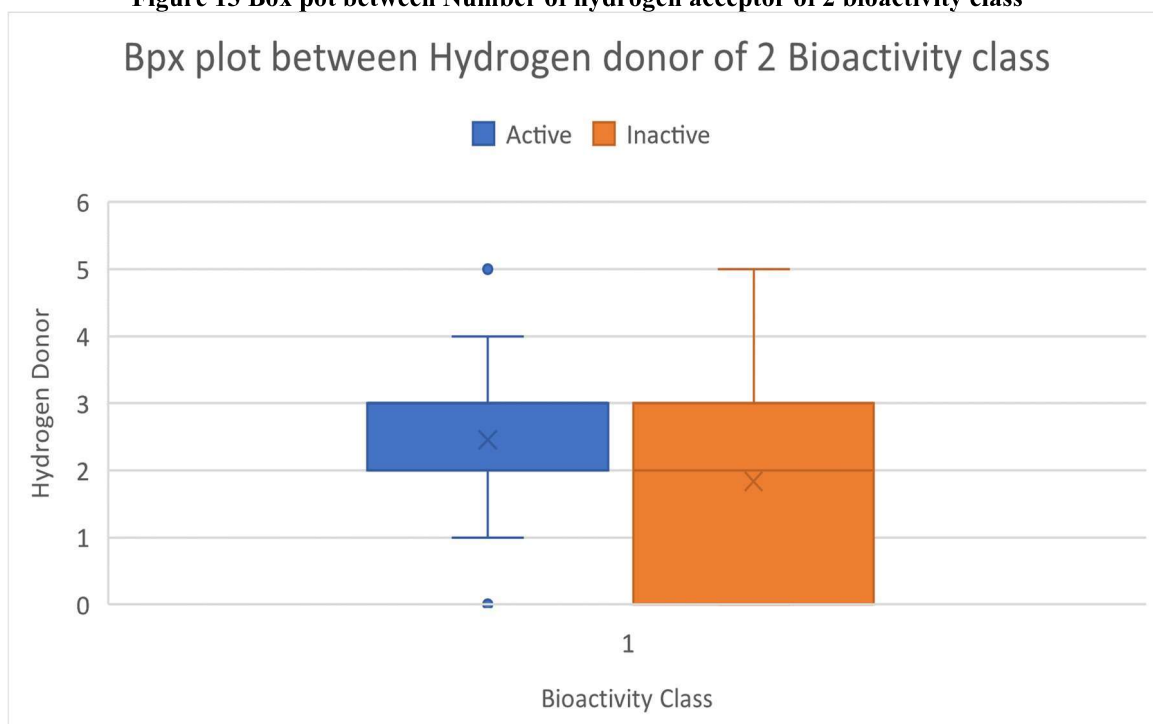


Figure 14 Box plot between number of hydrogen donor of 2 bioactivity class

9 Comparison of various regression model

Regression models help to perform various functions of relationship analysis between one or even more than on independent variables. The regression analysis yields a predicted value and performs various statistical analysis. The parameters which I have used to analyze model performance are R squared, Adjusted R squared and Root Mean Square Deviation (RMSE).

The preferred software is WEKA [36] for this. A data collection of 791 had been used in the QSAR model's construction. To transform the structural description into numerical values, various fingerprints were employed. Following data cleaning and filter selection, the dataset was divided into an 80/20 ratio, with 80% of the training data set and 20% of the test data set.



Figure 15 waikato environment for knowledge Analysis WEKA

REGRESSION MODEL	R- SQUARED	ADJUSTED R- SQUARE	RMSE
DECISIONTREEREgressor	0.89	0.92	0.32
EXTRA TREE REGRESSOR	0.86	0.9	0.32
RANDOM FOREST REGRESSOR	0.82	0.86	0.36
GAUSSIAN PROCESS REGRESSOR	0.80	0.85	0.36
KNEIGHBORS REGRESSOR	0.61	0.67	0.6
LINEAR SVR REGRESSION	0.48	0.66	0.69
LINEAR REGRESSION	-0.08	0.06	1.08
POISSON REGRESSOR	-0.12	0.00	1.08
SVR	-0.23	0.00	1.10
BAYESIAN RIDGE	-0.27	0.00	1.11

TABLE IV Showing various regression models accuracy on the basis of parameters R-squared, Adjusted-R Square, and RMSD.

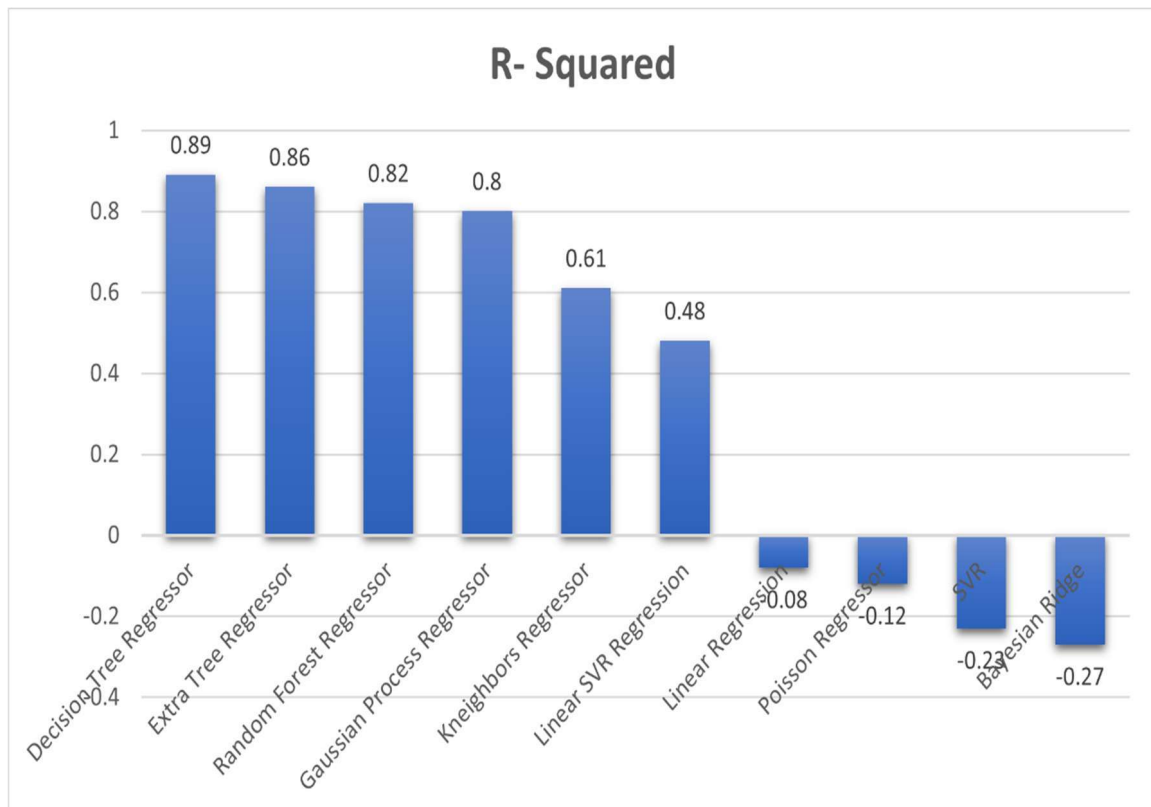


Figure 16 Comparison of Various regression models Vs R-squared

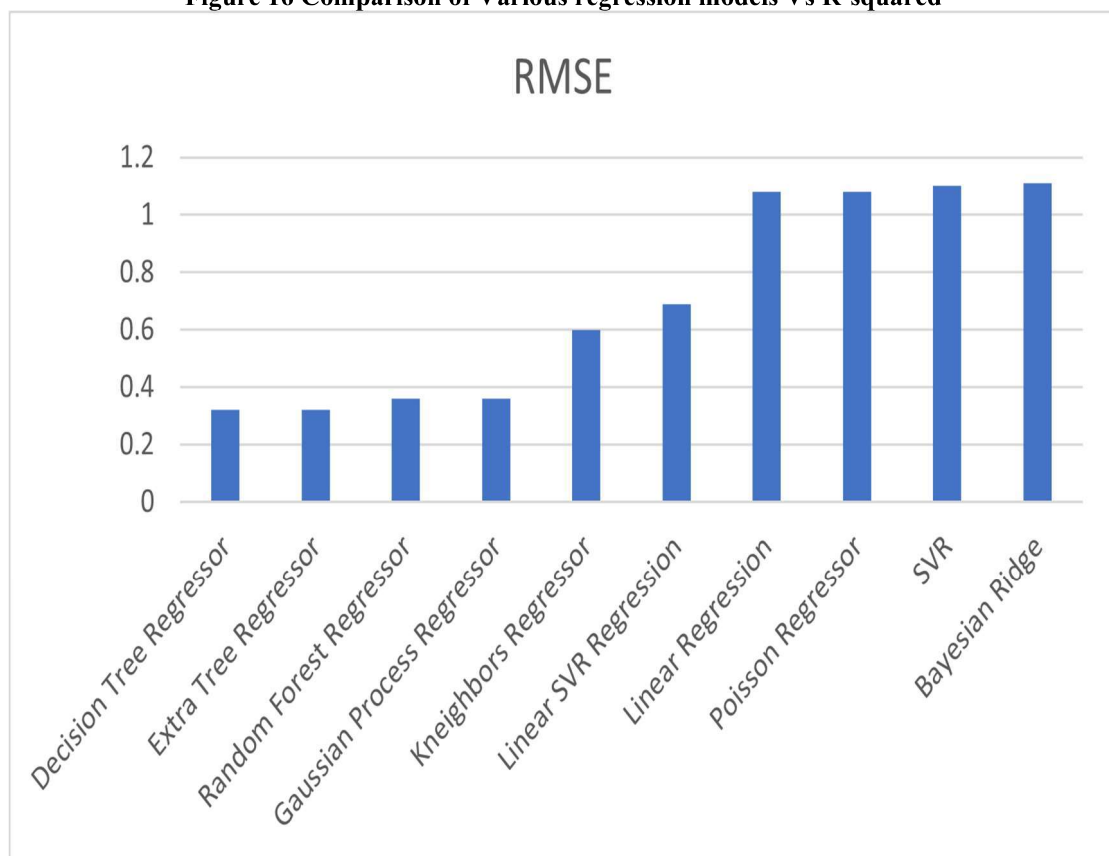


Figure 17 Comparison of Various regression model vs RMSE

10 Conclusion And Future Aspects

Despite various challenges in this synergy of AI in healthcare, there is going to be a promising future in the healthcare domain. Because of many success stories of this approach many researchers, healthcare experts, entrepreneurs are convinced and satisfied and believe that this synergy definitely assists the healthcare sector however the clinical workflow adaptation by AI approach and medical community concern make it challenging. This synergy needs approval by authorities and for this, the AI systems must have a trustworthy workflow and technology.

In the field of drug discovery and designing this tool make this easy and it lighten the burden on pharmaceutical domain. AI also impacted pathological approaches, nowadays pathologist have not to rely on microscopes for tissue examining they can study tissue now on their computer screen by using WSIs technology. Artificial intelligence synergy into pathology has also evolved way of interpretation, algorithms used by AI have ability to predict diagnostic result of a patient. US Food and Drug Administration (FDA) already approved AI application and AI in digital pathology could be next to be approved. In big data analysis the role of AI is not hidden it shows promising results in that field too. Efficient data mining techniques can open up a world of possibilities for data model analysis, revealing patterns that healthcare professionals can use in patient predictions, diagnosis, and treatment. One of the most serious problems of big data is the failure to preserve privacy, particularly in the case of data gathered from personal medical records. Despite the fact that there are standards in place to preserve the privacy of medical information, many of them do not apply to the transfer of big data. There are various up and low for all type of applications and Artificial intelligence is not the only one with

some issues. This technology has already evolved but needs some more efficient in order to deal with humans health. But despite of all AI will has promising result in healthcare and will definitely re-model the healthcare system.

11 References

- [1] J. H. Harrison *et al.*, “Introduction to artificial intelligence and machine learning for pathology,” *Arch. Pathol. Lab. Med.*, vol. 145, no. 10, pp. 1228–1254, 2021.
- [2] K. Bera, K. A. Schalper, D. L. Rimm, V. Velcheti, and A. Madabhushi, “Artificial intelligence in digital pathology - new tools for diagnosis and precision oncology,” *Nat. Rev. Clin. Oncol.*, vol. 16, no. 11, pp. 703–715, 2019.
- [3] B. X. Tran *et al.*, “Global evolution of research in Artificial Intelligence in health and medicine: A bibliometric study,” *J. Clin. Med.*, vol. 8, no. 3, p. 360, 2019.
- [4] “Artificial Intelligence (AI) in healthcare market – global drivers, opportunities, trends, and forecasts to 2022,” Kennethresearch.com. [Online]. Available: <https://www.kennethresearch.com/report-details/artificial-intelligence-ai-in-healthcare-market/10085225>. [Accessed: 25-Feb-2022].
- [5] T. Araújo *et al.*, “Classification of breast cancer histology images using Convolutional Neural Networks,” *PLoS One*, vol. 12, no. 6, p. e0177544, 2017.
- [6] J. Chen and C. Srinivas, “Automatic lymphocyte detection in H&E images with deep neural networks,” *arXiv [cs.CV]*, 2016.
- [7] P. Tschandl *et al.*, “Comparison of the accuracy of human readers versus machine-learning algorithms for pigmented skin lesion classification: an open, web-based, international, diagnostic study,” *Lancet Oncol.*, vol. 20, no. 7, pp. 938–947, 2019.
- [8] J. R. Naso *et al.*, “Discordance in PD-L1 scores on repeat testing of non-small cell lung carcinomas,” *Cancer Treat. Res. Commun.*, vol. 27, no. 100353, p. 100353, 2021.

- [9] S. Doyle, M. Feldman, J. Tomaszewski, and A. Madabhushi, "A boosted Bayesian multiresolution classifier for prostate cancer detection from digitized needle biopsies," *IEEE Trans. Biomed. Eng.*, vol. 59, no. 5, pp. 1205–1218, 2012.
- [10] A. Basavanthally *et al.*, "Multi-field-of-view framework for distinguishing tumor grade in ER+ breast cancer from entire histopathology slides," *IEEE Trans. Biomed. Eng.*, vol. 60, no. 8, pp. 2089–2099, 2013.
- [11] J. Barker, A. Hoogi, A. Depeursinge, and D. L. Rubin, "Automated classification of brain tumor type in whole-slide digital pathology images using local representative tiles," *Med. Image Anal.*, vol. 30, pp. 60–71, 2016.
- [12] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, pp. 533–536, 1986.
- [13] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [14] D. Bychkov *et al.*, "Deep learning based tissue analysis predicts outcome in colorectal cancer," *Sci. Rep.*, vol. 8, no. 1, 2018.
- [15] N. Coudray *et al.*, "Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning," *Nat. Med.*, vol. 24, no. 10, pp. 1559–1567, 2018.
- [16] G. P. Pena and J. S. Andrade-Filho, "How does a pathologist make a diagnosis?," *Arch. Pathol. Lab. Med.*, vol. 133, no. 1, pp. 124–132, 2009.
- [17] H. R. Tizhoosh and L. Pantanowitz, "Artificial intelligence and digital pathology: Challenges and opportunities," *J. Pathol. Inform.*, vol. 9, no. 1, p. 38, 2018.

- [18] X.-W. Chen and X. Lin, "Big data deep learning: Challenges and perspectives," IEEE Access, vol. 2, pp. 514–525, undefined 2014.
- [19] Office of the Commissioner, "FDA permits marketing of artificial intelligence-based device to detect certain diabetes-related eye problems," U.S. Food and Drug Administration, 11-Apr-2018. [Online]. Available: <https://www.fda.gov/news-events/press-announcements/fda-permits-marketing-artificial-intelligence-based-device-detect-certain-diabetes-related-eye>. [Accessed: 25-Feb-2022].
- [20] Office of the Commissioner, "FDA permits marketing of artificial intelligence algorithm for aiding providers in detecting wrist fractures," U.S. Food and Drug Administration, 24-May-2018. [Online]. Available: <https://www.fda.gov/news-events/press-announcements/fda-permits-marketing-artificial-intelligence-algorithm-aiding-providers-detecting-wrist-fractures>. [Accessed: 25-Feb-2022].
- [21] J. G. Elmore et al., "Diagnostic concordance among pathologists interpreting breast biopsy specimens," JAMA, vol. 313, no. 11, p. 1122, 2015.
- [22] D. V. M. Nicola M. Parry, "US pathologist supply down relative to diagnostic demands," Medscape, 31-May-2019. [Online]. Available: <https://www.medscape.com/viewarticle/913755>. [Accessed: 25-Feb-2022].
- [23] A. Esteva et al., "Dermatologist-level classification of skin cancer with deep neural networks," Nature, vol. 542, no. 7639, pp. 115–118, 2017.
- [24] Y. Jiang, M. Yang, S. Wang, X. Li, and Y. Sun, "Emerging role of deep learning-based artificial intelligence in tumor pathology," *Cancer Commun. (Lond.)*, vol. 40, no. 4, pp. 154–166, 2020.

- [25] T. Sakamoto *et al.*, “A narrative review of digital pathology and artificial intelligence: focusing on lung cancer,” *Transl. Lung Cancer Res.*, vol. 9, no. 5, pp. 2255–2276, 2020.
- [26] E. Rodner *et al.*, “Fully convolutional networks in multimodal nonlinear microscopy images for automated detection of head and neck carcinoma: Pilot study,” *Head Neck*, vol. 41, no. 1, pp. 116–121, 2019.
- [27] A. Kapil *et al.*, “Deep semi supervised generative learning for automated tumor proportion scoring on NSCLC tissue needle biopsies,” *Sci. Rep.*, vol. 8, no. 1, p. 17343, 2018.
- [28] Y. Guo, Z. Hao, S. Zhao, J. Gong, and F. Yang, “Artificial intelligence in health care: Bibliometric analysis,” *J. Med. Internet Res.*, vol. 22, no. 7, p. e18228, 2020.
- [29] J. Luo, M. Wu, D. Gopukumar, and Y. Zhao, “Big data application in biomedical research and health care: A literature review,” *Biomed. Inform. Insights*, vol. 8, pp. 1–10, 2016.
- [30] H. O. Alanazi, A. H. Abdullah, and K. N. Qureshi, “A critical review for developing accurate and dynamic predictive models using machine learning methods in medicine and health care,” *J. Med. Syst.*, vol. 41, no. 4, p. 69, 2017.
- [31] K. Y. Ngiam and I. W. Khor, “Big data and machine learning algorithms for health-care delivery,” *Lancet Oncol.*, vol. 20, no. 5, pp. e262–e273, 2019.
- [32] A. Bohr and K. Memarzadeh, “The rise of artificial intelligence in healthcare applications,” in *Artificial Intelligence in Healthcare*, Elsevier, 2020, pp. 25–60.
- [33] C. W. Yap, “PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints,” *J. Comput. Chem.*, vol. 32, no. 7, pp. 1466–1474, 2011.
- [34] M. Davies *et al.*, “ChEMBL web services: streamlining access to drug discovery data and utilities,” *Nucleic Acids Res.*, vol. 43, no. W1, pp. W612–20, 2015.

- [35] I. Ognjanovic, “Artificial intelligence in healthcare,” *Stud. Health Technol. Inform.*, vol. 274, pp. 189–205, 2020.
- [36] *Data mining: Practical machine learning tools and techniques*, 3rd ed. Oxford, England: Morgan Kaufmann, 2011.

PAPER NAME

Thesis copy.pdf

Ankit 2K20/MSCBIO/42

WORD COUNT

5915 Words

CHARACTER COUNT

36158 Characters

PAGE COUNT

38 Pages

FILE SIZE

3.2MB

SUBMISSION DATE

May 4, 2022 11:49 AM GMT+5:30

REPORT DATE

May 4, 2022 11:50 AM GMT+5:30

● 9% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

- 3% Internet database
- 1% Publications database
- Crossref database
- Crossref Posted Content database
- 8% Submitted Works database

● Excluded from Similarity Report

- Bibliographic material



● 9% Overall Similarity

Top sources found in the following databases:

- 3% Internet database
- Crossref database
- 8% Submitted Works database
- 1% Publications database
- Crossref Posted Content database

TOP SOURCES

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

1	Federal University of Technology on 2021-08-24	1%
	Submitted works	
2	Landmark University on 2021-08-17	<1%
	Submitted works	
3	Queen's University of Belfast on 2021-12-15	<1%
	Submitted works	
4	University of Westminster on 2021-11-25	<1%
	Submitted works	
5	CVC Nigeria Consortium on 2022-03-24	<1%
	Submitted works	
6	ebin.pub	<1%
	Internet	
7	University of Wolverhampton on 2021-02-10	<1%
	Submitted works	
8	fdic.gov	<1%
	Internet	

9	University of Hertfordshire on 2021-08-27	<1%
	Submitted works	
10	University of Surrey Roehampton on 2011-08-31	<1%
	Submitted works	
11	coursehero.com	<1%
	Internet	
12	Higher Education Commission Pakistan on 2021-06-19	<1%
	Submitted works	
13	University of Dallas on 2021-09-26	<1%
	Submitted works	
14	University of KwaZulu-Natal on 2022-02-10	<1%
	Submitted works	
15	University of Reading on 2020-01-13	<1%
	Submitted works	
16	etd.uwc.ac.za	<1%
	Internet	
17	Nazarbayev University on 2022-04-11	<1%
	Submitted works	
18	Queen Mary and Westfield College on 2022-04-06	<1%
	Submitted works	
19	Banaras Hindu University on 2020-02-19	<1%
	Submitted works	
20	bmcpharmacoltoxicol.biomedcentral.com	<1%
	Internet	

21	South Bank University on 2019-04-29	<1%
	Submitted works	
22	molecularbrain.biomedcentral.com	<1%
	Internet	
23	Bournemouth University on 2019-05-07	<1%
	Submitted works	
24	National Institute of Technology, Hamirpur on 2019-05-22	<1%
	Submitted works	
25	mdpi.com	<1%
	Internet	
26	Www.healthmeasures.net	<1%
	Internet	
27	Jihoon Jeong. "Deep Learning for Cancer Screening in Medical Imagin..."	<1%
	Crossref	
28	Nelson Marlborough Institute of Technology on 2016-07-15	<1%
	Submitted works	
29	Nottingham Trent University on 2020-09-13	<1%
	Submitted works	
30	The Robert Gordon University on 2020-04-06	<1%
	Submitted works	
31	link.springer.com	<1%
	Internet	
32	science.gov	<1%
	Internet	

-
- 33 Jonathan Phiri, Jackson Phiri, Charles S.. "Crime Mapping Model base... <1%
Crossref
-
- 34 University of East London on 2018-02-09 <1%
Submitted works

Artificial Intelligence and Digital Pathology Synergy

For detailed, accurate and predictive analysis of WSIs

Ankit

Department of Biotechnology
Delhi Technological University
Delhi, India
ankitchoudhary671998@gmail.com

Yasha Hasija*

Department of Biotechnology
Delhi Technological University
Delhi, India
Yashahasija06@gmail.com
*Corresponding author

Abstract— Pathology is the most important branch of Medical Science which fundamentally concerns the origin, cause, and nature of the disease. This domain contributes the most important part of the diagnostic infrastructure. The pathological examination includes examining, Tissues, Organs, Body fluids. In this examination, the tissue is fixed on a glass slide by certain procedures and then observed under a microscope for disease. Now with technological advancement, the traditional pathological operation is getting evolved and emerged as Digital pathology. Digital pathology comprises the digitalization of Histopathological slides using a Whole-slide scanner which is a microscope under robotic and computer control and analysing these digitalized images or whole-slide image using a Computational approach. This digitalization creates a high-resolution, enhanced pixel image of pathological specimen that ultimately generates a tremendous amount of data up to gigabytes per biopsy.

Because of algorithmic advancement and more convenient computing power, in the past few decades, Artificial Intelligence and machine learning have spread their roots into healthcare and clinical applications. And artificial intelligence and digital pathology synergy is one of the examples of these two fields intersection. This synergy will have majestic results in Diagnostic, prognostic, and predictive analysis of Whole-slide images.

Keywords— Artificial Intelligence, Deep Learning, Machine Learning, Whole Slide Images (WSIs), Histopathological Examination.

I. INTRODUCTION

Whole slide scanners form a digital image of histopathological slides and then AI technology is applied to these images for image processing and classification task in which the AI technology performs low-level tasks and high-level tasks respectively. In low-level tasks, object recognition problem is mainly focused like detection and segmentation and High-level tasks mainly focused on prediction of disease diagnosis and prognosis of disease by just analysing the digitalized pattern of the image. AI algorithms were designed in such a way that they can access and analyse the slide for certain tumors type and hence many AI algorithms need to design in order to get an accurate pathological report, in which tumors in tissue need to be diagnosed, graded, subtyped and accessing of high-

risk features should be done by another individual algorithm.

Various AI techniques (Fig.1.) machine learning, deep learning are used to convert WSI images into mineable data and these computation techniques come under Machine Vision. Machine vision is the technology that performs to extract information directly from an image using imaging-based analysis or inspection. By this approach interpretation of digital biopsies, samples could be done as quantitative datasets. Primary data performs a major role in AI performance. Quality of primary data decides AI performance. In order to achieve maximum prediction performance, primary data should have comprehensiveness, accuracy, and cleanliness.

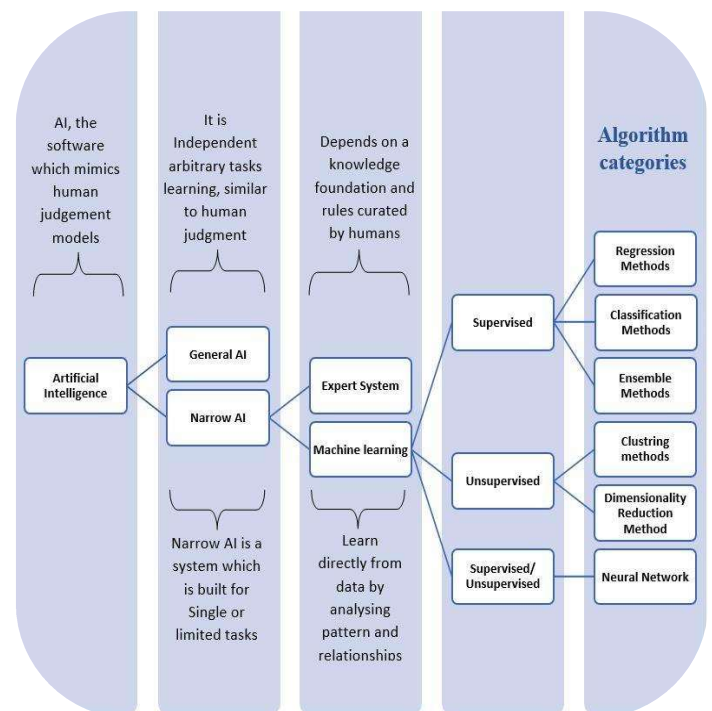


Fig. 1. Overview of different type of AI approaches and various algorithm categories.

II. AI PUBLICATIONS BOOM IN MODERN ERA

After technological advancements and the introduction of Digital Pathology, the market has grown remarkably with an exponential curve in the graph. The AI publication number drastically increased by more than 6 times within the last 20 years of the span. The Healthcare sector also came up with innovative ideas and starts utilizing or synergizing the computer science tool in their domain. Hence in result, the publications in biomedical literature have increased more than 8 times since 2000 and the Global AI market has estimated to be increased to \$300 billion from \$100 billion by 2025 and growth rate will be exponential with growth rate of around 40-50 % and Global AI in healthcare market is estimated to increase from \$856 million to \$20 Billion by the year 2025 [4]. Y Guo et al.[29] performed a bibliographic analysis of AI publications in healthcare. For this, they developed a strategy for searching published papers in the AI-related health care category. For this, the research papers including citations related to AI in healthcare data was retrieved from Web of Science (Wos). the publication records up to December 2019 were noted and a total of 5235 papers founded. From 5235 papers, 32 were excluded which are duplicates and currently proceeding, and 3730 were excluded because of not fulfilling screening criteria, so at last 1473 papers contributed to bibliometric analysis and steeply growth was seen from 2015-2019 with 987/1456 publications (67.78%). The timeline of the publications was given below (Table I) along with publications growth graph (Fig.2.)

TABLE I TIMELINE SHOWING THE NUMBER OF AI PUBLICATIONS AND ITS CONTRIBUTION PERCENTAGE IN THE AI-RELATED HEALTHCARE PUBLICATIONS.

YEAR OF PUBLICATION	NUMBER OF PUBLICATION	CONTRIBUTION PERCENTAGE %
1995-1999	38	2.60 %
2000-2004	97	6.66 %
2005-2009	98	6.73 %
2010-2014	236	16.20 %
2015-2019	987	67.78%
Total Publications	1456	100%

^a. Data source: - Data aggregated from study of Y Guo et al. [29]
<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7424481/>.

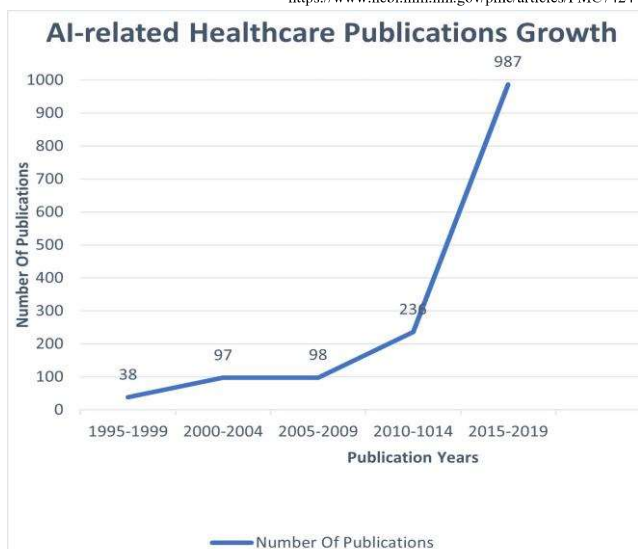


Fig. 2. AI-related Healthcare publications growth showing the rise in publications number with increasing year.

III. HISTOPATHOLOGICAL EXAMINATION USING AI

A. Using Deep Neural Network Approach

In this examination of whole slide images feature are directly extracted without explication of mathematical definitions. DL approach is mostly used and being adopted by digital pathology. DL easily learn representations directly from primary data and not necessarily depend on engineered features (Fig.3.). They are widely accepted because of high accuracy and easier application and hence preferred over hand-crafted feature engineering.

There are various Deep Learning algorithm have been developed like **Convolutional Neural Network (CNN)**; **Fully Convolutional Network (FCN)**; **Generative Adversarial Network (GAN)**; **Recurrent Neural Network (RNN)**.

1) Convolutional Neural Network (CNN)

Neural networks of these categories are used for Visual imaging analysis and because of this important feature they are the most enormously used DL algorithm for image analysis applications [5]. Based on their shared weight architecture of filters they are also called Space invariant artificial Neural Network (SIANN). The building block of CNN is convolutional sheets where the network learns and in between the input and output layer, there are filters that extract features of an image by analysing the grid-like topology. The image comprises various pixels patches which are arranged in a grid-like manner. CNN DL approach is used for Various tasks e.g., identification and quantification of cells (neutrophil, Lymphocytes, and blast cell). Tschandl et al. [7] organized a study that aims to compare the diagnostic accuracy between machine learning algorithms and expertise human readers. It is the open web-based diagnostic study in which 30 dermatoscopic image batches were selected randomly out of 1511 test set images has been asked to diagnose and then this diagnosis has been compared with 139 algorithms which had been created by 77 ML laboratories. The interpretation of this study goes in the favor of State-of-the-art machine-learning classifiers as they perform better than human readers. The DL algorithm has the more correct diagnosis and the mean of more correct diagnosis for the DL algorithm is 2.01 (17.91 vs 19.92; $P < 0.0001$) which is obtained by comparing human readers (n=511, human readers) against DL algorithms (n=139, algorithms), but they mentioned a limitation with algorithm performance because the performance of algorithm decreases for out-of-distribution images. Esteva et al. [24] demonstrated skin lesions classification by using a CNN approach. 129,450 clinical images were used by them which contains 2,032 different diseases that trains CNN model. The model is designed in such a way that it can differentiate malignant melanoma from benign nevi and keratinocyte carcinoma from benign seborrheic keratosis. The CNN model performance was compared with 21 board-certified dermatologists. N. Coudray et.al.[16] developed CNN

model named “inception v3” which classify Adenocarcinoma, Squamous cell carcinoma, and normal lung tissue accurately on WSIs which are provided by Cancer Genome Atlas. The performance of the model is founded with an average of 0.97 AUC.

2) Fully Convolutional Neural Network (FCN)

FCN model potentially detects scant quality pathological images by learning representations from every single pixel. It contains a convolutional layer hierarchy and in place of the last fully connected layer, it has a broad receptive region. Rodner et al.[27] developed FCN based image analysis algorithm. This algorithm can differentiate regions of malignancy from non-malignant regions of epithelium. This differentiation is based on non-linearity found in microscopic images of cancer of the head and neck. Head and neck cancerous sections were co-registered with multimodal images using microscopy. The WSIs are analysed and segmented into 4 categories malignant epithelium, Non-malignant epithelium, background, and other tissues. They analysed 114 WSIs which is obtained from 12 patients. The average recognition rate obtained is 88.9% and the overall recognition rate obtained is 86.7 of all of the four classes.

3) Generative adversarial network (GAN)

It is generative modeling that learns from regularities and patterns within the input data. In GAN two neural networks are in a competition with one another. simultaneously in a zero-sum game. One network is the “generator” and another one is the “discriminator”, Generator one generates

synthetic data while discriminator one compares the generated and original data for agreement. GAN is used by Gadermayr et al. who performs segmentation of glomeruli and segmented glomeruli out of WSIs from specimens of renal pathology which is isolated from resected mouse kidneys using GAN network. They selected mouse kidney because they are highly similar to human kidney. A. Kapil et al. [28] works on a deep learning model (GAN) for automated estimation of PDL1 and make the first automated scoring GAN model for PDL1 expression.

4) Recurrent Neural Network (RNN)

RNN is capable of storing input data on different time points for their sequential processing and learns from millions of discrete data [13]. RNN exhibits dynamic behavior as it stores input at a different time interval which is not seen in CNN and FCNs neural networks. CNN learns tasks by analyzing propagated errors and Long Short-Term Memory (LSTM) [14] plays a crucial role in this concern. LSTM is an RNN type that has recurrent gates or forget gates and these gates help CNN in task learning. A network which is been built by the combination of CNN and LSTM by Bychkov et al. [15]. This model analyses H&E-stained tissue of colorectal cancer Thrombotic microangiopathy (TMA) slide and predict disease re-occurrence. TMA images deconstructed small patches and then data is transferred into CNN. The small patches are easily understood and learned by the model. This combination

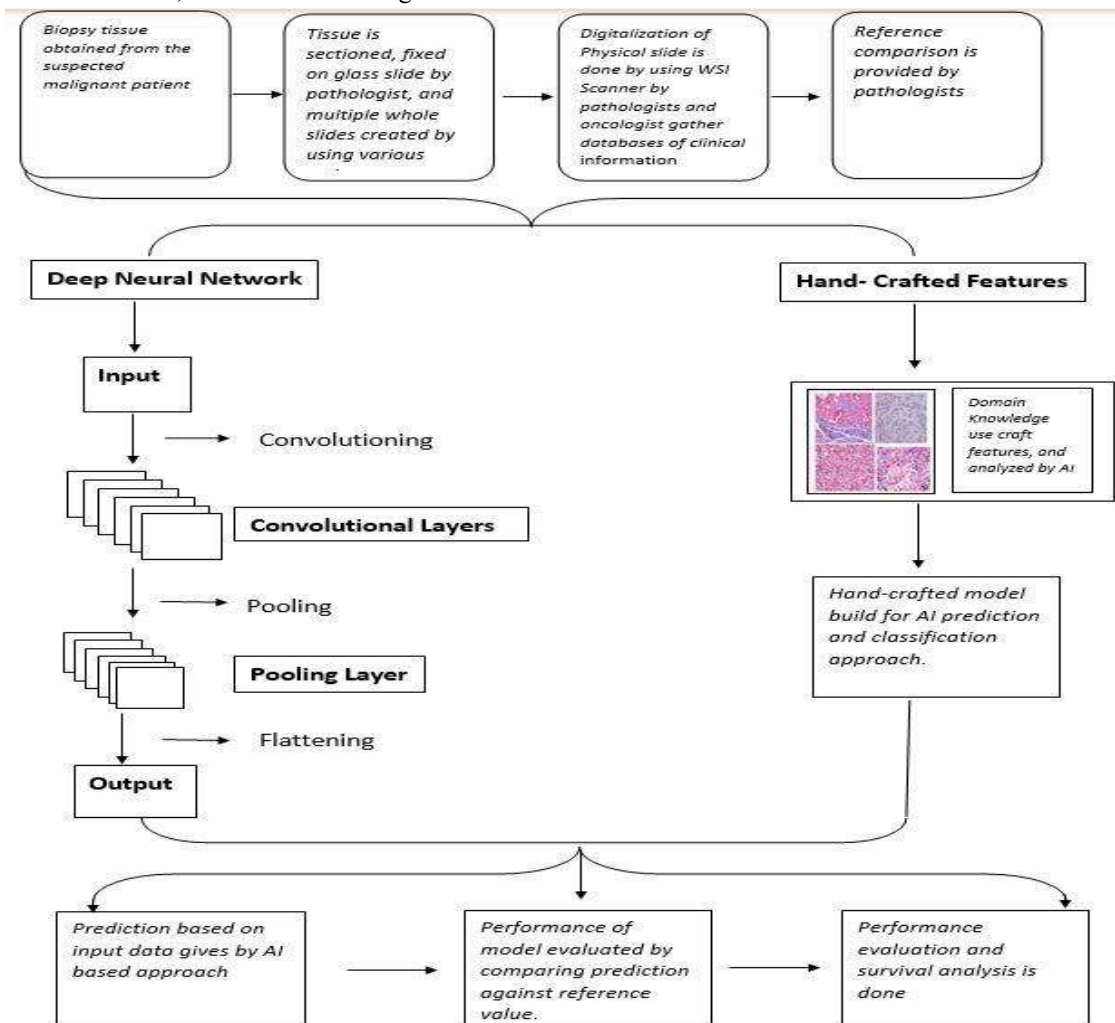


Fig. 3 Workflow of DNN and Hand-Crafted model in digital pathology [2]

(CNN+LSTM) gives the predictive performance of 0.69 AUC > 0.57 AUC of histological grade alone and the average score of visual risk (agreed by 3 pathologists) is 0.58 AUC [15].

TABLE II VARIOUS ADVANTAGES AND DISADVANTAGES OF DNN MODELS

DNN Model	Advantage	Disadvantage
Convolutional Neural Network (CNN)	High Accuracy; Automated extraction of features; fast to train	Large dataset requires; Explainability is low (CNN Black box)
(Fully Convolutional Network	Splits image of any size; shape, size and position of target is analyzed using pixel-level detection	Labelling cost is high
Generative Adversarial Network	Labeling isn't required; It learns and synthesizes realistic data	Training is complex and unstable.
Recurrent Neural Network	learns sequential Data (different time intervals)	Computational cost is high

B. Hand Crafted AI Approach

It is a traditional ML approach. In the handcrafted approach, the features and information present in an image have been used, and by applying a variety of algorithms the properties were derived. The algorithms are targeted towards specific features e.g., specific tissue type or cancer, and hence broad use is not a concern in this (Fig.3.). Simon et.al [10] have developed the automated system which is based on image imaging and could detect live prostate cancer cells. A Basavanahally et al. [11] developed a computerized system that distinguishes estrogen receptor-positive breast cancer tumor grade in a histopathological slide.

IV. WHY THIS SYNERGY IS IMPORTANT

According to a survey report by Medscape [23], the US is running out of pathologists and from 2007 to 2017 the number of pathologists were decreased by 17.53% and the workload on individual pathologists is raised by 41.73%.

This workload will affect the interpretations of pathologists and assistance of a technology (Artificial Intelligence) would definitely help in minimizing human errors.

There are various studies have been done on pathologists' discordance some of them are: -

A. Pathologists discordance in breast biopsy specimen interpretation

Pathology is a necessary component of the healthcare system, and nobody could neglect its role in diagnostics infrastructure. Pathologists ensure the quality of the report by accurately providing a quality report. But what if the quality of the report is not well as in some reports mentioning the discordance between pathologists for interpreting the same breast biopsy specimen. A group of

scientists (J. G. Elmore et al.) [22] designed an investigation to find out concordance among certain selected pathologists for interpretation of a given breast biopsy slide. The pathologist selection and their randomization criteria is given below (Fig.4.) The excisional or core needle breast biopsy specimen of briefly about 240 specimens were identified randomly by a 14,498 cases cohort obtained in Vermont and New Hampshire from pathology registries that are registered to Cancer Surveillance Consortium. There is a test set of 60 breast biopsies with 4 test sets giving around 240 total cases. Consensus panel gives consensus-derived reference diagnosis values which have 90.3% of concordance and these reference values of all 240 cases include 23 Invasive Breast Cancer cases, 72 Atypical Hyperplasia cases, 73 Ductal Carcinoma In Situ (DCIS) cases, and 72 Non-Benign Atypia cases.

For this investigation, pathologists are recruited publicly from 8 states of the US (Maine, Vermont, New Mexico, New Hampshire, Washington, Alaska, Minnesota, and Oregon).

B. Discordance in PDL1 tumor proportional scoring or TPS

PDL1 is a tumor biomarker, and its estimation is significant for the identification of Non-Small Cell Lung Cancer (NSCLC) in the patient. Currently, PDL1 expression quantification is done by visually estimating TPS scoring (percentage) of stained tumor cells. PDL1 expression has high spatial and temporal heterogeneity and this makes interpretation challenging. Julia R. Naso et al. [8] perform a study which is approved by the "University of British Columbia and BC Cancer Research Ethics Board". NSCLC patients were identified by retrospective chart review and in this PDL1 testing is done using 22C3 pharmDx assay which is a qualitative immunohistochemical assay that detects PDL1 tissue samples using Autostainer Link 48. Patients with two clinically requested PD-L1 tests were identified, and those tests had to be performed on the different pathological specimens and then from them, those patients were excluded who are suspect of two separated primary tumors. After that TPS scoring of tumor is clinically done by a pathologist group with repeat PDL1 testing and as result it is found that interpretation of 26 patients out of total of 77 patients has discordant PDL1 score with 36% of the discordant rate.

V. CHALLENGES AHEAD

This synergy of AI in pathology will not going to be that much easy as it seems. Many hurdles would be there in this path which we will have to beat. Accurate and comprehensive performance of AI is dependent on **Quality and Quantity of Input Data**, the input data must be clean along with high Signal : Noise ratio. The images been used for training must be labelled or annotated in order to delineate region of interest. This annotation work for larger number of images could be a challenge for pathologists and low-resolution images could make it even worse. The annotation issue could also be seen in Doyle et al. [9] work,

who developed Machine Learning based AI approach which detects prostate cancer region in WSIs. They found that when magnification of WSIs increases the performance

benign. The pathological diagnosis (pathological diagnosis has non-Boolean nature) action encompasses 4 domains cognitive, Communicative, Normative, Medical conduct

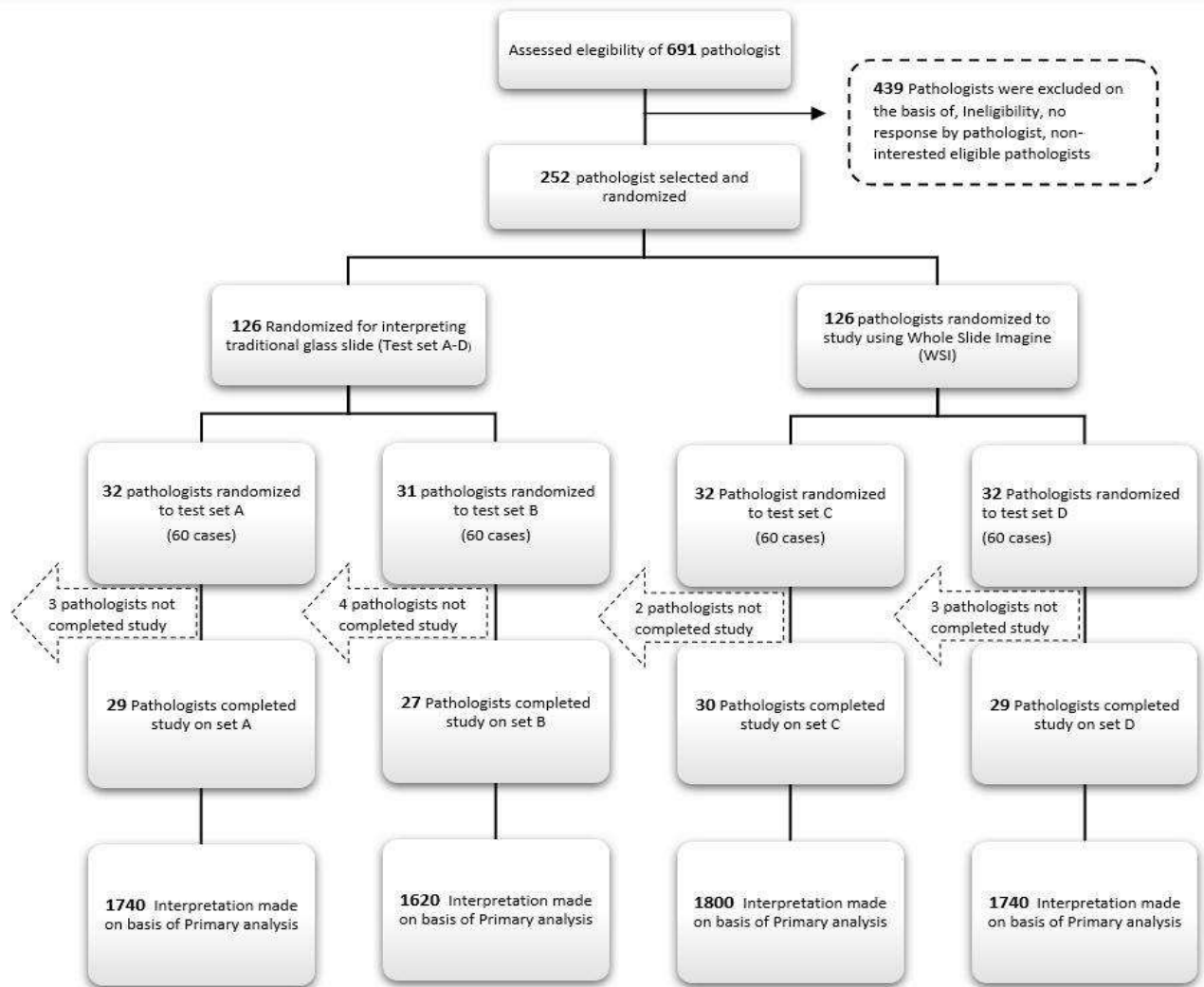


Fig. 4. Recruitment of pathologists and their randomization criteria [22]

decreases simultaneously and they concluded that this is due granularity loss and loss in reference learning manual annotations at high resolution. For the deployment of this technique AI tool is going to be app based which requires data cloud for instant sharing of pathological data. Challenges which are associated with this data cloud-based technique is that it requires larger bandwidth as size of WSI images is in gigapixel and hence there needs to be an uninterrupted channel between pathologists and cloud. Large datasets require enormous data and robust sharing and here **privacy consideration** would be obstacle [18]. In order to share health record of an individual there will be legal and technological hurdles. Binary variables which deal with digital pathology has only two possibilities of “yes” or “no” (**Boolean nature**) and diagnosis will be either malignant or

[17]. AI system requires fast processing of data and requires Graphical Processing Units (GPUs). This task is impractical in ordinary computers because CPUs are prohibitively sluggish [19] Hence being already in financial burdens for adopting WSI technology, **expenses affordability** is a major challenge for laboratories. There is continuous criticism which is been faced in the path regarding **lack of interpretability** despite of high accuracy of this method because the decision made by model has not any established path in other words there is not a verifiable path which enable us to understand the rationale behind decision hence remains unaccepted to medical community.

VI. CONCLUSION AND FUTURE ASPECTS

Despite various challenges in this synergy, there is going to be a promising future in the field of pathology. Because of many success stories of this approach many researchers, pathologists, entrepreneurs are convinced and satisfied and believe that this synergy definitely assists digital pathological tasks however the clinical workflow adaptation by AI approach and medical community concern make it challenging. This synergy needs approval by authorities and for this, the AI systems must have a trustworthy workflow and technology. In the field of radiology[21] and ophthalmology[20] US Food and Drug Administration (FDA) already approved AI application and AI in digital pathology could be next to be approved. The discordance in interpretation for various tasks by pathologists as mentioned above is a major concern for healthcare and the assistance of such a tool (Artificial Intelligence) would definitely assist diagnosis by providing an accurate report and shorter turnover rate. According to a survey report by Medscape [23], the US is running out of pathologists and from 2007 to 2017 the number of pathologists were decreased by 17.53% and the workload on individual pathologists is raised by 41.73%. AI adoption in digital pathology will decrease the burden on pathologists and diagnostic workflow will be effortless.

References

- [1] J. H. Harrison *et al.*, "Introduction to artificial intelligence and machine learning for pathology," *Arch. Pathol. Lab. Med.*, vol. 145, no. 10, pp. 1228–1254, 2021.
- [2] K. Bera, K. A. Schalper, D. L. Rimm, V. Velcheti, and A. Madabhushi, "Artificial intelligence in digital pathology - new tools for diagnosis and precision oncology," *Nat. Rev. Clin. Oncol.*, vol. 16, no. 11, pp. 703–715, 2019.
- [3] B. X. Tran *et al.*, "Global evolution of research in Artificial Intelligence in health and medicine: A bibliometric study," *J. Clin. Med.*, vol. 8, no. 3, p. 360, 2019.
- [4] "Artificial Intelligence (AI) in healthcare market – global drivers, opportunities, trends, and forecasts to 2022," Kennethresearch.com. [Online]. Available: <https://www.kennethresearch.com/report-details/artificial-intelligence-ai-in-healthcare-market/10085225>. [Accessed: 25-Feb-2022].
- [5] T. Araújo *et al.*, "Classification of breast cancer histology images using Convolutional Neural Networks," *PLoS One*, vol. 12, no. 6, p. e0177544, 2017.
- [6] J. Chen and C. Srinivas, "Automatic lymphocyte detection in H&E images with deep neural networks," *arXiv [cs.CV]*, 2016.
- [7] P. Tschandl *et al.*, "Comparison of the accuracy of human readers versus machine-learning algorithms for pigmented skin lesion classification: an open, web-based, international, diagnostic study," *Lancet Oncol.*, vol. 20, no. 7, pp. 938–947, 2019.
- [8] J. R. Naso *et al.*, "Discordance in PD-L1 scores on repeat testing of non-small cell lung carcinomas," *Cancer Treat. Res. Commun.*, vol. 27, no. 100353, p. 100353, 2021.
- [9] S. Doyle, M. Feldman, J. Tomaszewski, and A. Madabhushi, "A boosted Bayesian multiresolution classifier for prostate cancer detection from digitized needle biopsies," *IEEE Trans. Biomed. Eng.*, vol. 59, no. 5, pp. 1205–1218, 2012.
- [10] I. Simon, C. R. Pound, A. W. Partin, J. Q. Clemens, and W. A. Christens-Barry, "Automated image analysis system for detecting boundaries of live prostate cancer cells," *Cytometry*, vol. 31, no. 4, pp. 287–294, 1998.
- [11] A. Basavanahally *et al.*, "Multi-field-of-view framework for distinguishing tumor grade in ER+ breast cancer from entire histopathology slides," *IEEE Trans. Biomed. Eng.*, vol. 60, no. 8, pp. 2089–2099, 2013.
- [12] J. Barker, A. Hoogi, A. Depeursinge, and D. L. Rubin, "Automated classification of brain tumor type in whole-slide digital pathology images using local representative tiles," *Med. Image Anal.*, vol. 30, pp. 60–71, 2016.
- [13] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, pp. 533–536, 1986.
- [14] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [15] D. Bychkov *et al.*, "Deep learning based tissue analysis predicts outcome in colorectal cancer," *Sci. Rep.*, vol. 8, no. 1, 2018.
- [16] N. Coudray *et al.*, "Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning," *Nat. Med.*, vol. 24, no. 10, pp. 1559–1567, 2018.
- [17] G. P. Pena and J. S. Andrade-Filho, "How does a pathologist make a diagnosis?," *Arch. Pathol. Lab. Med.*, vol. 133, no. 1, pp. 124–132, 2009.
- [18] H. R. Tizhoosh and L. Pantanowitz, "Artificial intelligence and digital pathology: Challenges and opportunities," *J. Pathol. Inform.*, vol. 9, no. 1, p. 38, 2018.
- [19] X.-W. Chen and X. Lin, "Big data deep learning: Challenges and perspectives," *IEEE Access*, vol. 2, pp. 514–525, undefined 2014.
- [20] Office of the Commissioner, "FDA permits marketing of artificial intelligence-based device to detect certain diabetes-related eye problems," U.S. Food and Drug Administration, 11-Apr-2018. [Online]. Available: <https://www.fda.gov/news-events/press-announcements/fda-permits-marketing-artificial-intelligence-based-device-detect-certain-diabetes-related-eye>. [Accessed: 25-Feb-2022].
- [21] Office of the Commissioner, "FDA permits marketing of artificial intelligence algorithm for aiding providers in detecting wrist fractures," U.S. Food and Drug Administration, 24-May-2018. [Online]. Available: <https://www.fda.gov/news-events/press-announcements/fda-permits-marketing-artificial-intelligence-algorithm-aiding-providers-detecting-wrist-fractures>. [Accessed: 25-Feb-2022].
- [22] J. G. Elmore *et al.*, "Diagnostic concordance among pathologists interpreting breast biopsy specimens," *JAMA*, vol. 313, no. 11, p. 1122, 2015.
- [23] D. V. M. Nicola M. Parry, "US pathologist supply down relative to diagnostic demands," Medscape, 31-May-2019. [Online]. Available: <https://www.medscape.com/viewarticle/913755>. [Accessed: 25-Feb-2022].
- [24] A. Esteva *et al.*, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, pp. 115–118, 2017.
- [25] Y. Jiang, M. Yang, S. Wang, X. Li, and Y. Sun, "Emerging role of deep learning-based artificial intelligence in tumor pathology," *Cancer Commun. (Lond.)*, vol. 40, no. 4, pp. 154–166, 2020.
- [26] T. Sakamoto *et al.*, "A narrative review of digital pathology and artificial intelligence: focusing on lung cancer," *Transl. Lung Cancer Res.*, vol. 9, no. 5, pp. 2255–2276, 2020.
- [27] E. Rodner *et al.*, "Fully convolutional networks in multimodal nonlinear microscopy images for automated detection of head and neck carcinoma: Pilot study," *Head Neck*, vol. 41, no. 1, pp. 116–121, 2019.
- [28] A. Kapil *et al.*, "Deep semi supervised generative learning for automated tumor proportion scoring on NSCLC tissue needle biopsies," *Sci. Rep.*, vol. 8, no. 1, p. 17343, 2018.
- [29] Y. Guo, Z. Hao, S. Zhao, J. Gong, and F. Yang, "Artificial intelligence in health care: Bibliometric analysis," *J. Med. Internet Res.*, vol. 22, no. 7, p. e18228, 2020.

ACCEPTANCE MAIL

5/4/22, 2:13 PM

Gmail - IEEE - ICACCS 2022 - Acceptance Letter - Reg



Ankit Choudhary <ankitchoudhary671998@gmail.com>

IEEE - ICACCS 2022 - Acceptance Letter - Reg

1 message

ICACCS 2022 <icaccs@sece.ac.in>

Thu, Mar 3, 2022 at 1:15 PM

To: ankitchoudhary671998@gmail.com, yashahasija06@gmail.com

Dear Author(s),

Greetings from ICACCS 2022.

2022 8th International Conference on Advanced Computing and Communication Systems (ICACCS).

Organized by Sri Eshwar College of Engineering, Coimbatore, Tamilnadu, India.

We are very happy to inform you that your paper entitled "**Artificial Intelligence and Digital Pathology Synergy: For detailed, accurate and predictive analysis of WSIs**" with Paper-ID: **ICACCS_2022_paper_378** reviewed by the Technical Committee and recommended for presentation in the conference. The suggestions/comments made by the Technical Committee are listed at the end of this acceptance.

Kindly make all the necessary corrections as per their remarks before submitting camera ready paper. This is mandatory to include your paper in the proceedings. Formatting your paper with the below procedures and **keeping your paper without plagiarism, will increase the chance of publishing the paper in IEEE Digital Library publications.**

The "camera-ready" PDF paper can be created with IEEE PDF Express (**Conference-ID: 54159X**), (<https://ieee-pdf-express.org/>). All papers should strictly comply with IEEE Conference paper format attached with this mail and send the soft copy of camera ready paper both in PDF and WORD format to icaccs2022@gmail.com. **Kindly indicate your paper ID all the time in the subject line whenever you send mail to us.**

Kindly upload your Camera Ready Documents in icaccs.sece.ac.in -> Camera Ready Submission link.

The following guidelines are to be adhered to complete the registration process.

1. Before sending the PDF file generated through PDF express please ensure the format generated is in proper format and according to IEEE standard, otherwise try till proper format achieved. (Procedure is attached with this mail)

Note : IEEE Copyright notice will be updated by the conference publication Team, Authors no need to add the same

2. The Conference registration kindly visit our conference website icaccs.sece.ac.in

3. If the no. of pages of your paper exceeds 6 (Six), Rs.500 will be charged for an extra additional page, which should be included in your payment.

4. A Scanned copy of completed and signed **IEEE copyright form must be sent to icaccs2022@gmail.com** along with your Camera ready paper (**required for IEEE Digital Library Publication**).

5. All the accepted papers which are registered for IEEE ICACCS 2022 will be published in the IEEE conference proceedings with an ISBN number (assigned by IEEE, USA) and also submitted for inclusion in IEEE Digital Library, with author's concern.

You have to complete the registration process as per the steps given in our website icaccs.sece.ac.in

Review Reports

<https://icaccs.com/ICACCS2022/index.html>

Registration Link

<http://icaccs.sece.ac.in/Registration>

Camera Ready Link

<https://www.emailmeform.com/builder/form/MkflI3sm6ORWz>

General Editor Comments

Kindly verify the attached Plagiarism Report and improve your paper during Camera ready Submission.

We can accept plagiarism upto 15% during final submission. (only 3% Single source plagiarism recommended)

<https://mail.google.com/mail/u/1/?ik=eb703ad1ea&view=pt&search=all&permthid=thread-f%3A1726263882913861718&simpl=msg-f%3A172626...> 1/2

- o (Note: With reference 20%, without reference to 15% will be accepted)

Editor Comments for final Manuscript preparation :

1. Highlight the major contributions
2. Sections must be organized under the introduction.
3. Complete proofreading recommended
4. Image quality needs to improve.
5. Check the title and special characters
6. Remove unwanted/unused citations from your manuscript.
7. Add 6 to 10 keywords and organize sections under section 1.

Follow the proper IEEE Paper format.

****Detailed reviewers comments are available with the author terminal system/review report system.**

NOTE:

- (1) If you have more than one paper, you need to repeat the registration procedure.
- (2) If one of the authors registers for the paper presentation and if any of the other **co-authors** wish to participate, they should pay the participation fees.
- (3) The paid registration fee is non-refundable.
- (4) One regular registration is within Six Pages including all figures, tables, and references. Extra pages will be charged.
- (5) Filled registration form is one of the mandate requirements from the authors.

Authors' Registration Fee includes

1. Technical Sessions (Hybrid/Virtual/Online)
2. 10 Minutes oral presentation (Q&A included)
3. Conference Soft Proceedings. (Book format)
4. Softcopy of the Certificates.

Note:






- **IEEE Xplore accepts upto 20% plagiarism during your final submission; authors are requested to ensure the same during your camera ready submission. IEEE Recommend Cross Check / Turnitin for plagiarism verification.**
- **Registration Mapped with the first author mentioned in the paper. Authors need to pay the first author fee (Either Student or Faculty) during registration.**
- **Foreign (Indian) authors need to pay USD payments mentioned in the registration form.**

Any Queries Contact/ WhatsApp: 9486137910 or Email to: icaccs@sece.ac.in / anandakumar.h@sece.ac.in

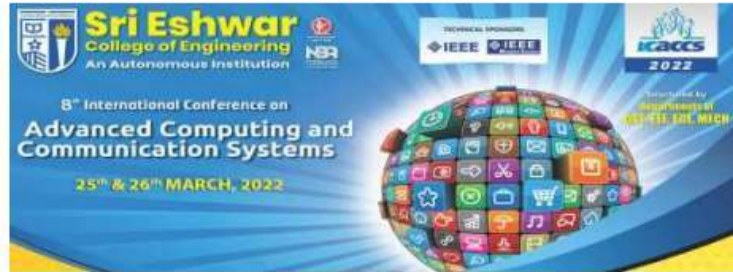
Thanks and Regards

**ICACCS 2022
Organizing Committee**

5 attachments

-  **54159X.PDF**
100K
-  **Acceptance Letter - ICACCS 2022.pdf**
364K
-  **ICACCS 2022 - IEEE - Copyright form.pdf**
106K
-  **IEEE_Paper_Format_ICACCS 2022.doc**
59K
-  **ICACCS_2022_paper_378.pdf**
2086K

ACCEPTANCE LETTER



Acceptance Letter

To
Dr./Ms./Mr. Ankit Ankit and Yasha Hasija

Paper ID: ICACCS_2022_paper_378

Dear Sir/Madam,

Sub: Acceptance Letter – IEEE 2022 8th International Conference on Advanced Computing and Communication Systems (ICACCS) 25th – 26th March 2022 Technically Sponsored by IEEE and IEEE Madras Section.

The organizing Committee is pleased to inform you that the peer- reviewed and refereed conference paper titled as “Artificial Intelligence and Digital Pathology Synergy: For detailed, accurate and predictive analysis of WSIs”, has been conditionally accepted for Hybrid (Oral/Virtual) presentation at the ICACCS 2022 conference on 25th – 26th March 2022.


We would like to kindly invite you to register for the conference on or before 10.03.2022 and present the paper at the conference venue in Coimbatore. On behalf of the organizing committee, I would like to congratulate you.

Note: Authors can present their paper through virtual / video conferencing.

Dr. H. Anandakumar
Conference Chair – ICACCS 2022


Sri Eshwar College of Engineering (Autonomous), Coimbatore, Tamil Nadu, India

Registration Fees Payment detail



[Share](#) [Tweet](#)

Your Registration for ICACCS 2022!



Organizer:
icaccs 2022


Hi ANKIT,

Thank you for purchasing 1 ticket(s) for **ICACCS 2022**.


Your Registration details are as follows:


Booking Id : **7667876**
Booking Date : **03/14/2022 22:12(IST)**


NAME	TYPE	PRICE
ANKIT	Indian Authors-Students(Non-IEEE Members)	INR 6500.00



Event Details :

 Organizer : **icaccs 2022**

 Venue : **Sri Eshwar College of Engineering, Coimbatore, India**

 Event Link : **icaccs2022**

(*Note : Please bring the printout of this email to the event OR show this on your smart phone at event venue)

Need any help?
Write to us at icaccs@sece.ac.in . We will get back to you shortly!

Create event of your own
You can also use townscript to manage registrations online!

[Create Event](#)

Certificate Of Presentation



ICACCS
2022

Event by & at :



Sri Eshwar
College of Engineering
An Autonomous Institution
Coimbatore, India



2022
8th International Conference on
Advanced Computing & Communication Systems

TECHNICAL SPONSORS





Certificate of Presentation

Certify that

Ankit

Delhi Technological University, Delhi, India.

has presented a paper in the International Conference on
Advanced Computing & Communication Systems - ICACCS 2022
on 25th & 26th March 2022 at Sri Eshwar College of Engineering,
Coimbatore, TamilNadu, India.

Paper Title :
Artificial Intelligence and Digital Pathology Synergy



Dr. H. Anandakumar
Conference Chair



Dr. R. Subha
Convener



Dr. Sudha Mohanram
Patron



ICACCS