APPLICATION OF ML TO MAKE SENSE OF BIOLOGICAL BIG DATA IN DRUG DISCOVERY PROCESS

A PROJECT REPORT

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE AWARD OF THE DEGREE

OF

Master of Science

In

Biotechnology

Submitted by:

Divya Sharma

2K19/MSCBIO/11

Under the supervision of:

DR. YASHA HASIJA

Assistant Professor

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

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

CANDIDATE'S DECLARATION

I Divya Sharma, Roll Number: 2K19/MSCBIO/11, student of M.Sc. Biotechnology, hereby declare that the work which is presented in the Major Project entitled "**Application of ML to make sense of Biological Big Data for Drug Discovery**" 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 2021, under the supervision of Dr. 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: Big Data: A Boom in biomedical sciences
Author Names: Sharma, Divya and Hasija, Yasha
Name of Conference: 12th International Conference on Computing, Communication and Networking Technologies (ICCCNT) – IEEE Conference
Conference Date and Venue: 6th-8th July 2021 at Indian Institute of Technology Kharagpur (IIT Kharagpur or IIT-KGP), Kolkata, West Bengal
Registration: Done
Status of Paper: Acceptance Received
Date of Paper Communication: 30th April 2021
Date of Paper Acceptance: 28th May 2021
Date of Paper Publication: NA

Date: 29th May 2021 Divya Sharma

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

CERTIFICATE

To the best of my knowledge, the above work entitled "**Application of ML to make sense of Biological Big Data for Drug Discovery**" 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, has not been submitted in part or full for any Degree or Diploma to this University or elsewhere. I, further certify that the publication and indexing information given by the student is correct.

Place: Delhi

Date: 29th May 2021

31-05-2021

Dr. Yasha Hasija
(Supervisor)
Assistant Professor
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, Dr. 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 to carry out the research and to present my work works as clearly as possible. It was a great privilege and honor to work and study under her guidance. I am extremely grateful for what he has offered me. Her insightful feedback pushed me to sharpen my thinking and brought my work to a higher level.

I am extremely grateful to my parents for their love, prayers, caring and sacrifices for educating and preparing me for my future.

I would also like the institution Delhi Technological University, Delhi for giving me the opportunities throughout the tenure of study.

Finally, my thanks go to all the people who have supported me to complete the research work directly or indirectly

Divya Sharma

Abstract

Scientists have been working over years to assemble and accumulate data from biological sources to find solutions for many principal questions. Since a tremendous amount of data has been collected over the past and still increasing at an exponential rate, hence it now becomes unachievable for a human being alone to handle or analyze this data. Most of the data collection and maintenance is now done in digitalized format and hence requires an organization to have better data management and analysis to convert the vast data resource into insights to achieve their objectives. The continuous explosion of information both from biomedical and healthcare sources calls for urgent solutions. Healthcare data needs to be closely combined with biomedical research data to make it more effective in providing personalized medicine and better treatment procedures. Therefore, big data analytics would help in integrating large data sets for proper management, decision-making, and cost-effectiveness in any medical/healthcare organization. The scope of the thesis is to highlight the need for big data analytics in healthcare, explain data processing pipeline, and machine learning used to analyze big data.

Contents

LIST OF FIGURES

List of Tables

Chapter 1

Introduction

Data assimilation and its subsequent analysis are integral parts of the better functioning of every organization. There have been several advances in the field of medical science and computation which led to a noticeable merging of the healthcare and the life science sectors that eventually enhance the relationships between the patients, their doctors, and biopharmaceutical companies.

To keep up with the increasing pace of the dynamically growing field of medicine, organizations all over the world are researching and proposing many different healthcare information system models to provide the best services and care to their patients. These are machine learning models, majorly dependent on data produced by electronic health records as well as the complex omics and biomedical data.

Due to the increase in new data sets, it is necessary to evaluate the qualitative standards of the unstructured data and use data mining techniques to derive proper insights out of it which could help physicians and researchers to understand how to use the data for improving healthcare services. Big data analytics has now started to be used in clinical practices leading to offering personalized and precision healthcare to the people.

Drug discovery is a process which is designated approach to find out a molecule for detailed evaluation as a potential drug candidate [1] for curing and treating a disease. It is a time involving process and requires enormous amounts of money to be invested by the pharmaceutical companies [2] [3]. The possible drug candidates undergo a series of examination procedures to be qualified for further development and clinical trials. It takes about 12-15 years for a novel drug candidate to reach the market after undergoing all the required procedures [4]. Every drug target research involved thousands of compounds out of which eventually one qualifies approval. It is very necessary to understand the fundamental process in the drug discovery pipeline in order to comprehend the difficulties of achieving a single medication to the patients [5].

Availability of health and biomedical big data does provide several unexpected opportunities but also brings along challenges concerning data analysis and data mining [6]. The techniques like ML and DL benefit from the enormous amount of data available on the public biological database platforms like ChEMBL [10] and PubChem [9]. The technologies are mainly used to differentiate the molecules having slightly similar properties while predicting their biological activities and establishing correlations between them.

From the view of scientific discovery, using this extensive amount of information for drug discovery would not only increase the horizon of knowledge but also would help predict numerous hypotheses to make the disease prediction much more reliable. It would eventually decrease the time spent over browsing the existing literature leading to the process becoming faster [7]

Considering the new health challenges in the population, there is always a need of newer drugs in the market. Discovery of a drug involves a various processes involving target identification and validation, hit identification, lead generation and optimization and eventually the identification of a candidate for further development. All these processed require time and efforts along with the need to meet the validation requirements from the regulatory bodies. [8] [9] Here, I present the use of the knowledge of big data and machine learning to help in reducing the time and efforts involved in the R&D processes.

Chapter 2

Big Data Analysis for Healthcare

The word "big data" implies accumulating complex and large data sets, which have computational, storage, and communication potential that cannot be met by traditional methods or systems [10].

The successful completion of the Human Genome Sequencing led to the generation of big biomedical data. The evolving high throughput and omic technologies are significantly contributing to the increasing volume of biological and biomedical data [11]. Big data techniques can be used to design and formulate newer architecture to derive insights from massive volumes of varied data involving a high-velocity capture and analysis [12]. The change in computing architecture would enable the researchers to handle the heavy processing which is required to analyze the huge heap of data in a secured way [13]. To make the best use of big data, companies use real-time information for creating newer products and services by identifying patterns in the data. This information could be utilized to provide effective treatments and cures for life-threatening and even rare diseases [14].

2.1. Big Data Characteristics

Big Data is centered on the below mentioned characteristics.

2.1.1 Volume

The volume of data generated in any organization is increasing at an exponential rate. Numerous issues related to storage and analysis are associated with increasing data [15] [16].

Due to the lack of proper tools for processing a large amount of data, a large amount of it is overlooked, deleted, or delayed from processing [17].

2.1.2. Velocity

The speed at which the information is being produced, treated, and transferred is increasing at an accelerating rate [15]. The biomedical data can be generated in real-time or in batches and contribute to the rapidity of the manufactured data. With the increased involvement of smart medical devices and newer technologies like the Internet of Things, data traffic has increased at a pace higher than data consumption and processing [18] [19].

2.1.3. Variety

The data received is in a complex structured, semi-structured, or structured form [15] which is incompatible due to structural differences. It is varied in the matter of source, data types, and information contained. The variety and richness of the data play an indispensable role in developing the big data strategy [20].

2.1.4. Value

The sole aim of using BDA on healthcare and biomedical data is to obtain valuable insights from it and to provide better deliverables in terms of services. Better analysis leads to smarter and intelligent decisions by creating maximum value from all the total volume of the data that is produced in the industry with each passing day [21].

2.1.5. Veracity

Many times the quality of data is at a compromise due to various noise factors associated with it causing it to be uncertain and less operational [18].

Variability, Validity, Vulnerability, Volatility, and Visualization are equally important characteristics. These dimensions are very important for the issues that arise in big data analytics [22].

2.2. Big Data Processing

The tools required for the analysis of biomedical and healthcare BDA are different from those used for traditional data analytics. The huge amount of data produced demands for tools with high complexity, requiring intensive programming skills. The information and insights extracted can thereby open new doors of knowledge and transformation in an innovative manner [23].

The process of BDA for healthcare data can be explained in the following steps.

2.2.1. Data Acquisition

The entire healthcare industry is expected to see a change because of the transition in the data from normal to big data in regards to size, diversity, and complexity [18]. With the advent of technology, there has been a change in data storage practices from traditional physical file- based system to more accessible computational record systems [24].

There is an increased amount of biomedical data available from the medical history of different patients, diagnosis details, prescriptions, medical images, [24] many government-sponsored projects, repositories, pharma R&D projects, and various research based investigations [11].

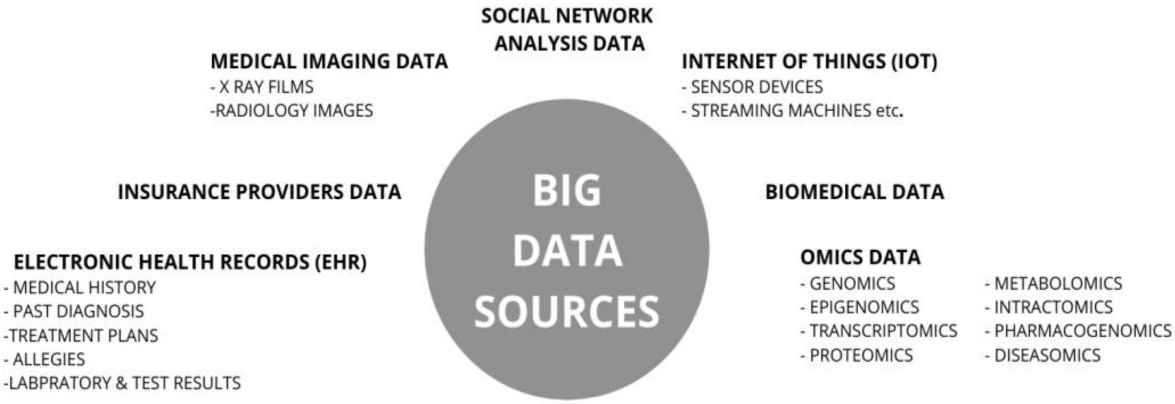


Fig 1: Some Common Sources for Healthcare Big Data

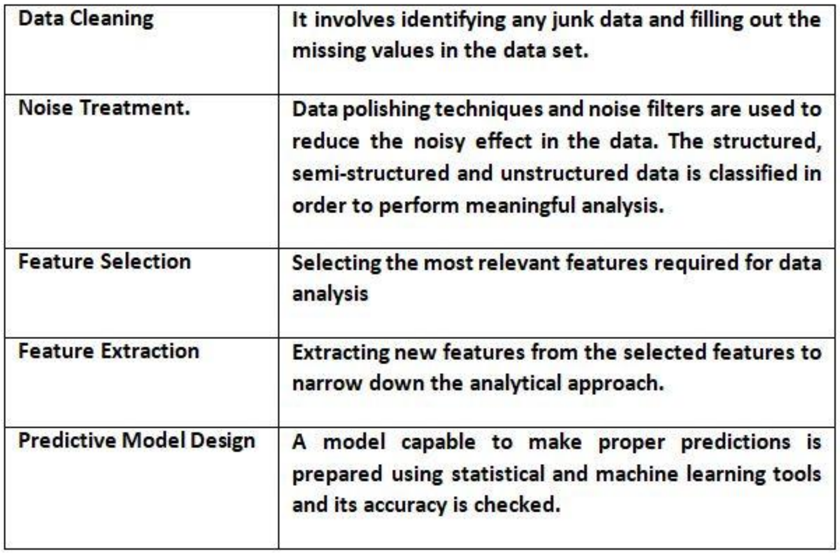
2.2.2 Data Storage

Even after advancements in both computation and internet services, there is a void in the infrastructure which is required in proper storage of the data before and after analysis. The storage of data has now become more expensive than it would have cost to produce the same amount of data. Cloud computing can be the only storage method that provides the desired elasticity for the storage of healthcare Big Data. Different companies provide different platforms that can be used as Dropbox for the data for both storage and transfer [16] [25] [26].

2.2.3. Data Management

The efficient management of data involves cleaning data, retrieving necessary information from unstructured data, and data mining [27]. The cleaning of data involves processing the data for noise reduction and managing the missing values. Medical records contain lots of unwanted information which is required to be filtered out to avoid discrepancies in data [28]. Proper Big Data management would help to increase the dependability towards the healthcare procedures which results in real-time information and to make more precise predictions depending on the condition of the patient.

Table 1: Data Management Steps



Data Cleaning	It involves identifying any junk data and filling out the missing values in the data set.
Noise Treatment.	Data polishing techniques and noise filters are used to reduce the noisy effect in the data. The structured, semi-structured and unstructured data is classified in order to perform meaningful analysis.
Feature Selection	Selecting the most relevant features required for data analysis
Feature Extraction	Extracting new features from the selected features to narrow down the analytical approach.
Predictive Model Design	A model capable to make proper predictions is prepared using statistical and machine learning tools and its accuracy is checked.

2.2.4 Data Analytics

There is a need for proper analysis and modification of the generated data to enhance the standards of the services provided, also to ensure proper coordination for patients. The unprocessed data is redundant and is useless. The analysis should aim in administering beneficial decisions by identifying patterns and relationships amongst diverse data sets. This is commonly done by using various machine learning algorithms to bring a revolution to the current clinical research [22]. The data needs to be clearly labeled to train the machine learning algorithms and model building [29]. There are two data mining techniques that are used to analyze data based on classification, regression, cluster analysis, text mining etc. [30].

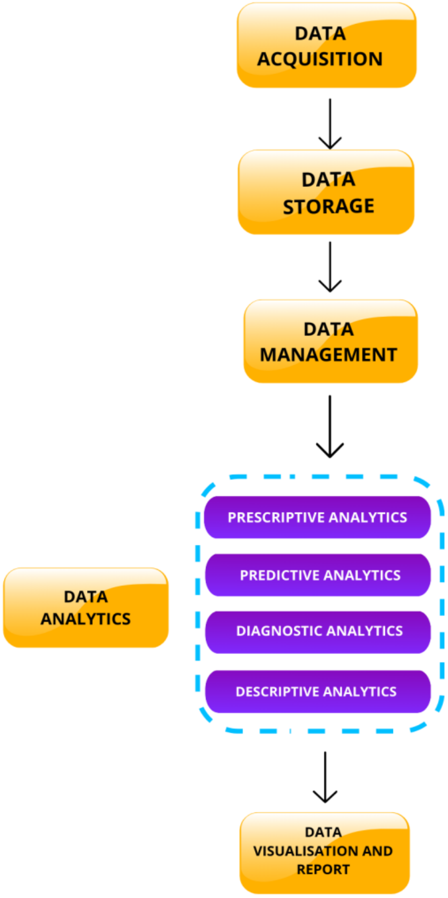


Fig 2: Data processing life cycle

Descriptive: Various events of past are explained in details and data mining is done on that entire data to obtain detailed information about the occurrence that had taken place sometime in the past. Such analysis can provide information about the beginning of the spread of the disease, cause, diagnosis, treatments that were carried out. It would also provide insights about the kind and combination of the drugs that were prescribed as well as the lifestyle changes that were recommended [18]. It is a form of unsupervised learning.

Diagnostic: The events of the descriptive analytics are used to identify the different causes of those events on order to find the reasons for the disease outbreaks. It helps to jump onto conclusions about why certain diseases are dormant over the others and in what conditions they could become active again. Computational techniques like data extraction, data mining and data correlation are used to derive insights from the past data [18].

It accesses the association between the different prognostic factors and healthcare outcomes [31].

Predictive: Using data mining, statistical modeling, ML and AI techniques on the existing data, future events can be predicted. Such an analytical method is called Predictive Analysis. It helps to analyze the prevailing health conditions and could help preventing adverse health conditions [10], thus pose the potential to minimize health risk. Big data technologies are used to discover useful patterns from the data that would help in prediction leading to cost reduction [32].

Prescriptive: It includes the ability to suggest the best decisions or options based on the prescriptive analysis output. The prescriptive analytics based algorithms have enabled the health professionals and doctors to be able to provide patients with timely prescriptions and to treat various complications. Prescriptive model lifecycle involves model building, model solving and model adapting [33]. It is the complement to predictive analytics. Problem solving in prescriptive analytics categorized in 4 distinct categories - Blind Search, Local Search, Search based on Population and Multi-objective Optimization [34].

2.2.5. Data Visualization

Visualization of data has an integral contribution in providing the results of the extensive analysis that has been performed on the data. It concludes all the steps starting from the collection, cleaning to analyzing of data. Proper visualization of data is essential for a better understanding of medical professionals. Data can be presented in form of charts, dashboards, maps, tables, diagrams, etc [35]**.**

Various tools that can be used to analyze the data include – R language, Graphviz, Google Charts, Tableau, iCharts, SAS Visual Analytics, etc. [16].

2.3. Big Data in Biological Science

Bioscience and medicine have generated massive amounts of data, much of which is freely available for analysis. Researchers with access to such information can investigate and comprehend the mechanisms which might cause diseased states, and also the potential to diagnose and treat them. Publicly supported organisations and institutions have been established to function as data contributors to make such resources more accessible to the scientific community. The European Bioinformatics Institute (EMBL-EBI) and the National Center for Biotechnology Information (NCBI) are two examples.

The completion of the Human Genome Sequencing resulted in the collection of great amount of biomedical data. When multiple resources are combined and presented together, there is a considerably better potential to study and analyse data. The ultimate aim of any data-driven method in the biomedical science is to gain insights that help enhance global human health.

Applications for Big Data in drug discovery research and development range from clinical study design to understanding how to target biological systems to influence disease processes. While there is a lot of promise in using Big Data to improve clinical results and patient care, failure to consider privacy and data protection issues can lead to violations of the law and a loss of public trust in the institutions.

Chapter 3

Machine Learning for Healthcare

Doctors need quick and precise predictions for diagnosis of their patients' disorders to take better treatment decisions. With the help of machine learning, intelligent models have been developed for data collection, analysis, storage, and usage [36].

Machine learning facilitates the understanding of hidden patterns in data by using existing algorithms on big datasets. Various statistical and mathematical methods are used to extract knowledge from these datasets [29]. The ML algorithms can be categorized into supervised and unsupervised depending upon the kind of input-output used to train the ML model. The input for training the model can have labeled supervised) or unlabelled (unsupervised) examples. Unsupervised and supervised learning can be used consecutively to discover hidden messages within the data [6].

It is important to understand the scope and intent before building the right model for analysis. Clustering and classifier models are the most commonly used predictive models in decision making. Specific subsets of machine learning i.e. neural networks or deep learning models are trained to accurately find patterns in the highly complex big data. Deep learning involves many variables and several input-output layers which are missing in the traditional analysis/statistical methods. These multilayer neural networks help establish complex relationships among the variables to create prediction models for handling big data. Other machine learning models such as SVM, decision trees, etc have been able to prove effective in addressing healthcare and biomedical data [29][36].

Efficiently used ML algorithms can help the doctors and physicians to derive conclusions for accurate diagnosis, prescribe the best medications to the patients, upgrade the patient's general health standards and help to identify the patients who are susceptible to repeated illness [37]. Thus, it is very essential to understand the limitations of the model as well as to elucidate the inferences.

3.1. Algorithms used in machine learning

There are four types of machine learning algorithms: supervised, unsupervised, semi-supervised, and reinforcement learning.

3.1.1. Supervised learning

Algorithms that require external aid are known as supervised machine learning algorithms. The training and testing datasets are separated from the input dataset. The output variable from the train dataset is that needs to be predicted or categorized. For prediction or classification problems, the algorithms learn patterns through the training dataset and apply the learning on the test dataset. [38].The most used supervised learning algorithms are mentioned below

3.1.1.1. Decision Trees

Decision trees are trees that group features by ordering them according to their values. The decision tree is primarily used for classification-problems. Nodes and branches make up each tree. Each branch indicates a value that the node can take, and each node represents attributes in a group that needs to be categorized [38].

Decision trees are a dependable and effective decision-making method that uses high classification accuracy with a simple representation of obtained knowledge. They've been employed in a variety of medical decision-making applications [39].

3.1.1.2. Naïve Bayes

The text categorization industry is the prime aim of Nave Bayes. It is mostly employed for the purposes of clustering and classification [40]. The conditional probability is used in the underlying architecture of Nave Bayes. It constructs trees depending on the likelihood of them occurring. Bayesian Network is another name for these trees.

Naive Bayes classifiers are probability based classifiers that apply Bayes' theorem to features with high (naive) independence assumptions. The model is easy to build and does not need iterative parameter estimate, that makes it particularly useful in the field of medicine. [41] The Bayes classifier can evidentially achieve the best result provided the probability distribution. The Naive Bayesian Classification approach has also been used to develop decision support in the Heart Disease Prediction System. Data mining techniques may aid in solving various significant and vital problems linked to health care by storing and digitalizing treatment records of millions of patients. The finest decision support system is the Nave Bayes classification. [42]

3.1.1.3. Support Vector Machine

It is another popular ML approach that is primarily employed for categorization. SVM is predicated on the concept of calculating margins. Basically, it is used to construct margins between the different classes. The margins are set so that the gap between the margin and the classes is maximized, reducing the classification error [43].

3.1.2. Unsupervised learning

The data is just used to train a few features via the unsupervised learning methods. When new data is introduced, it recognizes the data's class using previously learnt features. It's mostly utilized for feature reduction and clustering.

3.1.2.1. K-Means Clustering

This algorithm facilitates in data grouping, with K being the group number. Iteratively allocates each data point to a group which is dependent on the features provided. The similarity of the feature is then used to cluster the data points. The centroids of the K clusters and labels for the data are the results of K-means clustering.

3.1.3. Semi-Supervised learning

Semi-supervised learning algorithms combine the benefits of both supervised and unsupervised learning techniques. It can be useful in fields like machine learning and deep learning if there is existing unlabeled data and collecting the labelled data is a time-consuming procedure [44]. In the field of medical science the semi-supervised learning is extensively used for medical image classification [45].

3.1.4. Reinforcement learning

Reinforcement learning is a sort of learning in which the learner makes choices about which steps to perform in order to improve the outcome. Until a situation is presented, the learner has no understanding of what actions should be taken. The learner's activities may have an impact on events and their behaviour in the future situations. Reinforcement learning is based primarily on two concepts: trial and error searching and delayed results. [46]

To assist medical practitioners and patients in intervening at an earlier stage, pre-analyze illnesses and therapies. It also aids in the identification of public health hazards by recognizing patterns, model disease progression etc. [47]

3.2. Applications of Machine Learning

3.2.1. In Bioinformatics: Data management systems [48] like MapReduce and Hadoop are the most widely used in the bioinformatics field to serve better data warehouse repositories, infrastructure for computing and data mining tools to analyze biological information in manageable time frame [49]. Both open access and commercial tools are now used for clinical genomic analysis to unravel the sequestered messages in the genomic data. However, the use of these tools is challenging in terms of the expertise and experience to create value out of the information. The organizations have to choose from the other tools available in the market, the right tool that would solve their business problems [48].

3.2.2. In Medicine: Individual medical datasets have been incorporated into big data algorithms to encourage the use of evidence-based treatment to provide better and an open-information era in the healthcare industry [48]. A careful examination of the clinical data, patient's records, genomic data is essential to determine the need of new treatment procedures [23].

3.2.3. In Remote Monitoring: The use of sensory smart devices has encouraged both caregivers and the people to engage efficiently. This has led to conscientious monitoring of the patient by keeping a continuous check on their condition. Additionally, the increased use of **Internet of Things** (IoT)—enabled devices among the patients helps reduce the visits to the hospitals, reducing the overall treatment costs and prevents re-admissions. [23] [49].

3.2.4. In Research and Development: The use of predictive modeling encourages the technocrats to come up newer and more efficient models for extracting information. Establishing connection between the data from diverse sources unravels information to discover adverse effects of the disease, diagnosis, treatment, drug or even a medical device before-hand, saving on major costs [23].

3.2.5. In Public Health: A common pattern of diseases or co morbidities can be identified among specific groups of people and most accurate treatment can be provided to them accordingly. More absolute treatment can be given by preparing targeted medicines and vaccines by carefully determining the requirements of the recipients [23].

3.2.6. In Pharmaceuticals: Big Data is useful in all the stages of development of pharmaceuticals, especially in making of precision medicine. Using the EHR data collected for analysis, conditions of the patients can be closely analyzed to deliver most accurate form of treatment [16].

3.2.7. In Medical Imaging: The reports generated using deep learning techniques help the doctors to analyze the life-threatening and cancerous diseases better than the earlier times [21].

Chapter 4

Machine learning in Drug Discovery and Design

One of the most significant and constantly expanding domains in computer-aided drug development is machine learning [50]. Machine learning approaches are much more effective than physical models and therefore can be scaled to large datasets without requiring a lot of processing power. One of the most common uses of machine learning in drug development is to aid researchers in understanding and exploiting correlations between chemical compositions and biological activity, often known as SAR [51].

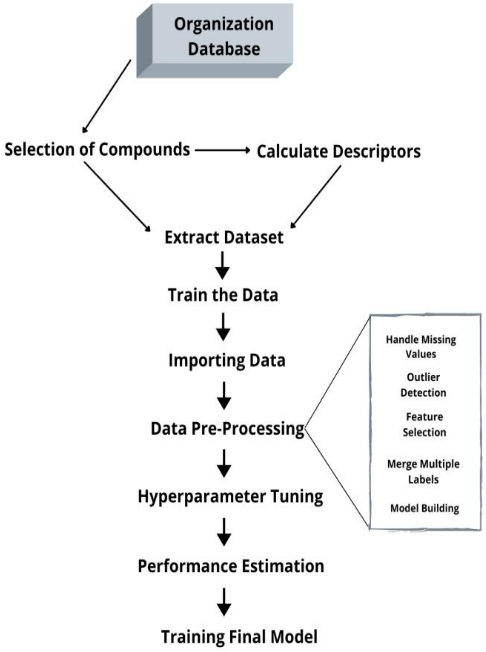


Fig 3: An example work flow for building machine learning models from raw data for drug research and development.

Machine learning and statistical techniques have been used to find the biological activities of the compounds that are of interest in the drug development. Various quantitative structure activity relationship (QSAR) models are developed for understanding bioactivity of the compounds by implementing machine learning algorithms to the molecular descriptors.

For faster and effective studies, there was a requirement of conversion of the molecular structural information into one or more numerical values for establishing quantitative correlations amongst structures and characteristics, biological activities, or various other experimental properties of the compounds. Molecular descriptors are such numerical representations of the compounds generated by applying a well-defined algorithm on defined molecular description or a well-defined experimental process for a specific compound. They serve a significant importance in the researches involving quantitative structure–activity relationships (QSARs) [52].

The QSAR models have been majorly using the regression models to relate the logIC50 values with the structural and chemical characteristics of the different compounds under study. [53]. Advancement of technology has led to the conversion of the chemical structure to chemically informatics that can be applied ML algorithms which use the input containing molecular descriptors as fingerprints to generate correlations between the different biological activity parameters under study.

4.1. Literature Overview

Psoriasis can be defined as the most common chronic, inflammatory immune-mediated skin disorder affecting almost 1-115 of the population worldwide [54].

90% of Psoriasis cases can be considered as chronic plaques (Psoriasis vulgaris). A number of identified factors such as Trauma, Stress and pathogens are found to trigger Psoriasis which affects the extensor surfaces of the limbs; scalp and the trunk. Pustular Psoriasis can progress rapidly and is characterized by generating multiple generalized or local pustules with diffused redness. Inflammation in the joints can lead to Psoriatic arthritis. Psoriatic inflammation is caused due to aggregation of immune cells, most commonly dendritic cells and macrophages which ultimately leads to extensive cytokine signalling and Ɣ𝜶T-cell immune activation.

Type 1 interferons (IFNƔ and IFN𝜶) are responsible for the maturation and activation of dendritic cells which driving towards T-cell regulated inflammatory responses. Macrophages and Dendritic cells promotes the release of TNF-𝜶 which is further responsible for triggering the circuit of inflammatory interleukin cascade (IL-17; IL-21; IL-22 etc.). The TNF-𝜶 - IL-23- Th-17 is responsible for T-cell mediated Psoriasis is one of the major players involved in this inflammatory mediated disorder.

A thorough investigation of literature reviews briefs us about the importance of TNF-𝜶 as one of the important targets for restricting the progression and advancement of Psoriasis. Multiple studies suggest that the use of monoclonal antibodies against TNF has been able to reduce the inflammation and interleukin responses in Psoriasis patients.

TNF-α shRNA treatment showed decreased levels of TNF-α mRNA as detected in skin biopsies 3 weeks after a single vector injection of lentiviral vectors encoding TNF-α shRNA [55]. The results are consistent with the hypothesis that increased IFN-γ and TNF-α in psoriasis is associated with a systemic pro-inflammatory gradient in the skin, which then promotes inflammatory responses in both aortic endothelial cells [56]. TNF-α targeted therapies, as well as new molecules and compounds targeting TNF-α, will continue to play an important role in the lifelong management of psoriasis [57].

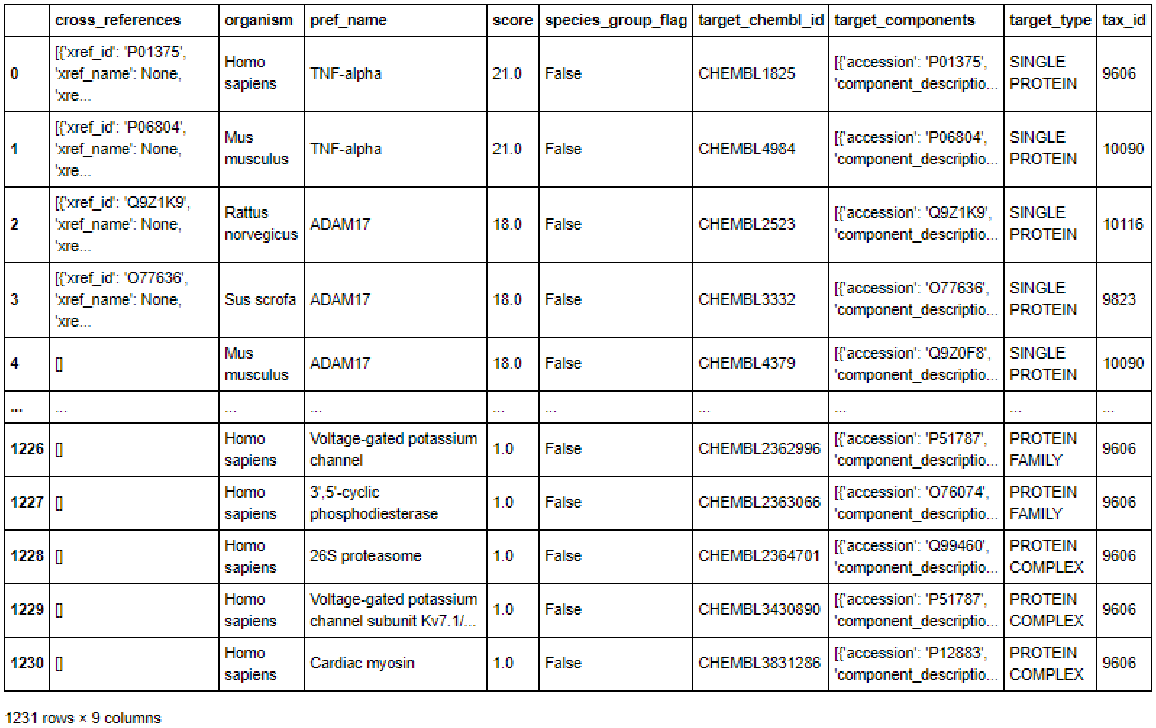
Here, ML based regression methods have been employed to understand the biological activity associated with TNF- α inhibition which is responsible for the progression of psoriasis. The structure based approach has been used in order to unravel the desired molecules pertaining to the drug discovery.

4.2. Methodology

Machine learning techniques essentially depend upon the quality of the dataset. Various biological activity data providing databases are available online that provide data comprising of quanititative information IC50, EC50, Ki and potency. The summary of the workflow is provided in the figure. It involves the following steps:

4.1. Data Set Preparation

4.2.1.1. Extraction of bioactivity data from ChEMBL Database: To find all probable target compounds, a search is conducted in the CHEMBL database for TNF-alpha. ChEMBL (https://www.ebi.ac.uk/chembl) is a database of bioactive compounds with drug-like characteristics that has been manually curated. It combines chemical, bioactivity, and genetic data to aid in the translation of genetic data into novel medications that work. The dataset obtained contains 1231 targets. Bioactivity data is retrieved for TNF-alpha for Homo sapiens using the unique **ChEMBL ID “CHEMBL1825**”. Bioactivity data is an essential form of scientific data that must be searchable, accessible, interoperable, and reusable [58]. Pharmacological/biological activity is one of the most important qualities of chemical compounds since it indicates how the chemicals might be used in medical applications.



	cross_references	organism	pref_name	score	species_group_flag	target_chembl_id	target_components	target_type	tax_id
0	[{'xref_id': 'P01375', 'xref_name': None, 'xre...	Homo sapiens	TNF-alpha	21.0	False	CHEMBL1825	[{'accession': 'P01375', 'component_descriptio...	SINGLE PROTEIN	9606
1	[{'xref_id': 'P06804', 'xref_name': None, 'xre...	Mus musculus	TNF-alpha	21.0	False	CHEMBL4984	[{'accession': 'P06804', 'component_descriptio...	SINGLE PROTEIN	10090
2	[{'xref_id': 'Q9Z1K9', 'xref_name': None, 'xre...	Rattus norvegicus	ADAM17	18.0	False	CHEMBL2523	[{'accession': 'Q9Z1K9', 'component_descriptio...	SINGLE PROTEIN	10116
3	[{'xref_id': 'O77636', 'xref_name': None, 'xre...	Sus scrofa	ADAM17	18.0	False	CHEMBL3332	[{'accession': 'O77636', 'component_descriptio...	SINGLE PROTEIN	9823
4	[]	Mus musculus	ADAM17	18.0	False	CHEMBL4379	[{'accession': 'Q9Z0F8', 'component_descriptio...	SINGLE PROTEIN	10090
...	...	...	...	...	...	...	...	...	...
1226	[]	Homo sapiens	Voltage-gated potassium channel	1.0	False	CHEMBL2362996	[{'accession': 'P51787', 'component_descriptio...	PROTEIN FAMILY	9606
1227	[]	Homo sapiens	3',5'-cyclic phosphodiesterase	1.0	False	CHEMBL2363066	[{'accession': 'O76074', 'component_descriptio...	PROTEIN FAMILY	9606
1228	[]	Homo sapiens	26S proteasome	1.0	False	CHEMBL2364701	[{'accession': 'Q99460', 'component_descriptio...	PROTEIN COMPLEX	9606
1229	[]	Homo sapiens	Voltage-gated potassium channel subunit Kv7.1/...	1.0	False	CHEMBL3430890	[{'accession': 'P51787', 'component_descriptio...	PROTEIN COMPLEX	9606
1230	[]	Homo sapiens	Cardiac myosin	1.0	False	CHEMBL3831286	[{'accession': 'P12883', 'component_descriptio...	PROTEIN COMPLEX	9606

1231 rows × 9 columns

Fig 4: Search results for TNF-α from ChEMBL Database

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

Out of the several bioactivity measurement units IC50 was chosen for further analysis because they made up a major subset of compounds. An inhibitor's IC50 is the concentration at which response (or binding) is halved. The standard value indicates the potency of the drug; the lower the value, the more potent the medicine. The number of the standard value should be as low as possible for an idealistic condition, i.e. the inhibitory concentration at 50% should be low. This implies that a lower drug concentration would be required to achieve 50% inhibition of the target protein. Therefore, only those values were retrieved for which are reported as IC50 values in nM (nanomolar) unit. Other compounds were removed because they either did not have any values for IC50 or lesser/greater than the required values.

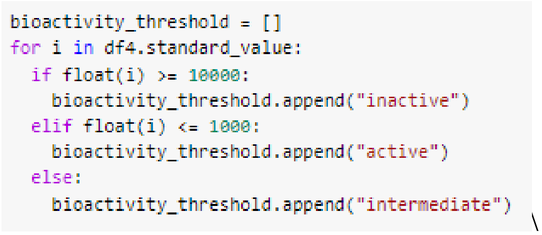
4.2.1.3.Handling missing data and removing duplicates: The dataframe is checked for any missing values in the **standard_value** and **canonical_smiles** column.The missing values are dropped before labeling of the molecules is done. This was done to make the data more uniform for further analysis. A quality dataset with 959 compounds was obtained. There were many compounds having same values for the canonical smiles notation. These redundant values were removed in order to have only unique values in the final dataset.

4.2.1.4. Labeling of the compounds: The chemicals are classified as active, inactive, or intermediate based on their IC50 values.

(i) The compounds having IC50 values >= 10,000nm are classified as inactive class,

(ii) The compounds having values <= 1000nm are classified as active class, and

(iii) The compounds having values in between to that of 1000 and 10000 were considered as intermediate class



```
bioactivity_threshold = []
for i in df4.standard_value:
  if float(i) >= 10000:
    bioactivity_threshold.append("inactive")
  elif float(i) <= 1000:
    bioactivity_threshold.append("active")
  else:
    bioactivity_threshold.append("intermediate")
```

\

Fig 5: Python Code for classifying the classes based on their threshold IC50 values

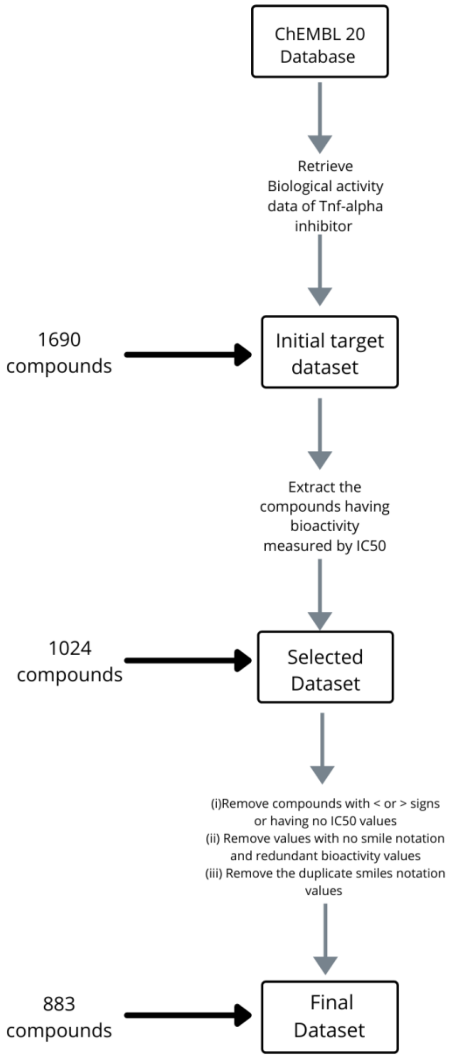
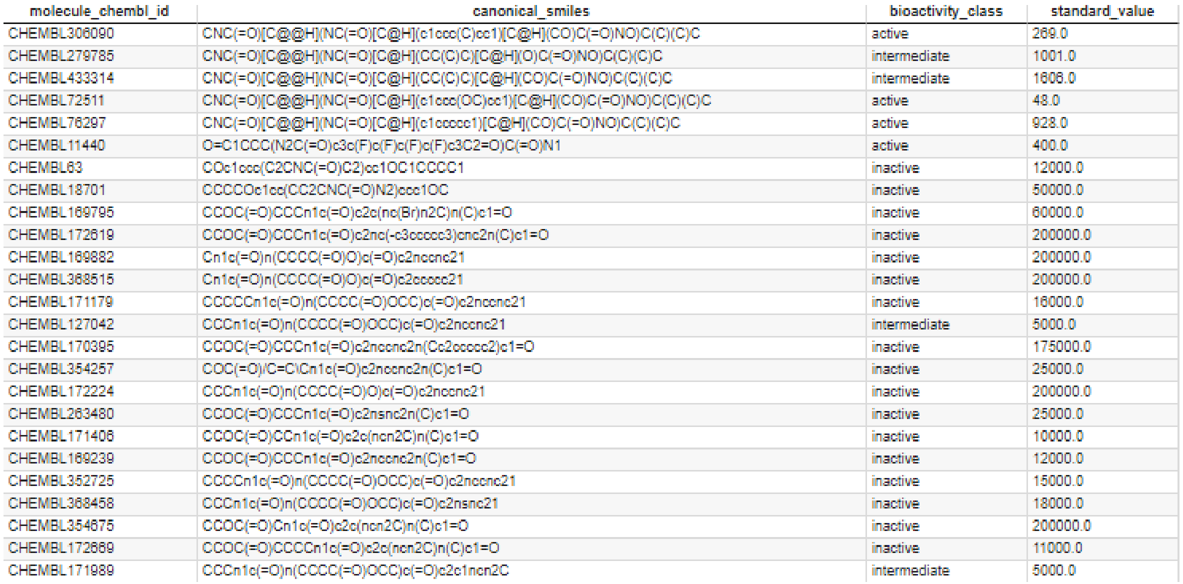

ChEMBL 20 Database
Retrieve Biological activity data of Tnf-alpha inhibitor
1690 compounds
Initial target dataset
Extract the compounds having bioactivity measured by IC50
1024 compounds
Selected Dataset
(i)Remove compounds with < or > signs or having no IC50 values
(ii) Remove values with no smile notation and redundant bioactivity values
(iii) Remove the duplicate smiles notation values
883 compounds
Final Dataset

Fig 6: General scheme for dataset preparation for Exploratory Data Analysis

4.2.2. Data Pre-Processing

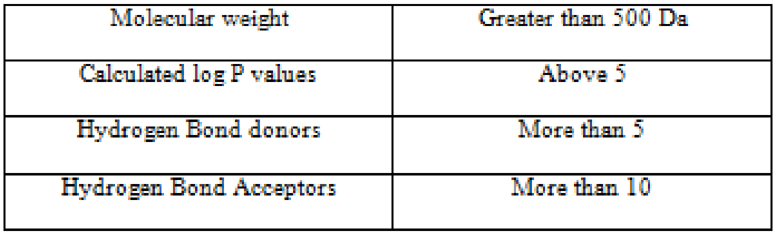
4.2.2.1. **Feature selection:** A new data subset was created containing columns molecule_chembl_id, canonical_smiles, standard_value and bioactivity_class. It was further used for data pre-processing steps.



molecule_chembl_id	canonical_smiles	bioactivity_class	standard_value
CHEMBL306090	CNC(=O)[C@@H](NC(=O)[C@H](c1ccc(C)cc1)[C@H](CO)C(=O)NO)C(C)(C)C	active	269.0
CHEMBL279785	CNC(=O)[C@@H](NC(=O)[C@H](CC(C)C)[C@H](O)C(=O)NO)C(C)(C)C	intermediate	1001.0
CHEMBL433314	CNC(=O)[C@@H](NC(=O)[C@H](CC(C)C)[C@H](CO)C(=O)NO)C(C)(C)C	intermediate	1606.0
CHEMBL72511	CNC(=O)[C@@H](NC(=O)[C@H](c1ccc(OC)cc1)[C@H](CO)C(=O)NO)C(C)(C)C	active	48.0
CHEMBL76297	CNC(=O)[C@@H](NC(=O)[C@H](c1ccccc1)[C@H](CO)C(=O)NO)C(C)(C)C	active	928.0
CHEMBL11440	O=C1CCC(N2C(=O)c3c(F)c(F)c(F)c(F)c3C2=O)C(=O)N1	active	400.0
CHEMBL63	COc1ccc(C2CNC(=O)C2)cc1OC1CCCC1	inactive	12000.0
CHEMBL18701	CCCCOc1cc(CC2CNC(=O)N2)ccc1OC	inactive	50000.0
CHEMBL169795	CCOC(=O)CCCn1c(=O)c2c(nc(Br)n2C)n(C)c1=O	inactive	60000.0
CHEMBL172619	CCOC(=O)CCCn1c(=O)c2nc(-c3ccccc3)cnc2n(C)c1=O	inactive	200000.0
CHEMBL169882	Cn1c(=O)n(CCCC(=O)O)c(=O)c2nccnc21	inactive	200000.0
CHEMBL368515	Cn1c(=O)n(CCCC(=O)O)c(=O)c2ccccc21	inactive	200000.0
CHEMBL171179	CCCCCn1c(=O)n(CCCC(=O)OCC)c(=O)c2nccnc21	inactive	16000.0
CHEMBL127042	CCCn1c(=O)n(CCCC(=O)OCC)c(=O)c2nccnc21	intermediate	5000.0
CHEMBL170395	CCOC(=O)CCCn1c(=O)c2nccnc2n(Cc2ccccc2)c1=O	inactive	175000.0
CHEMBL354257	COC(=O)/C=C\Cn1c(=O)c2nccnc2n(C)c1=O	inactive	25000.0
CHEMBL172224	CCCn1c(=O)n(CCCC(=O)O)c(=O)c2nccnc21	inactive	200000.0
CHEMBL263480	CCOC(=O)CCCn1c(=O)c2nsnc2n(C)c1=O	inactive	25000.0
CHEMBL171406	CCOC(=O)CCn1c(=O)c2c(ncn2C)n(C)c1=O	inactive	10000.0
CHEMBL169239	CCOC(=O)CCCn1c(=O)c2nccnc2n(C)c1=O	inactive	12000.0
CHEMBL352725	CCCCn1c(=O)n(CCCC(=O)OCC)c(=O)c2nccnc21	inactive	15000.0
CHEMBL368458	CCCn1c(=O)n(CCCC(=O)OCC)c(=O)c2nsnc21	inactive	18000.0
CHEMBL354675	CCOC(=O)Cn1c(=O)c2c(ncn2C)n(C)c1=O	inactive	200000.0
CHEMBL172669	CCOC(=O)CCCCn1c(=O)c2c(ncn2C)n(C)c1=O	inactive	11000.0
CHEMBL171989	CCCn1c(=O)n(CCCC(=O)OCC)c(=O)c2c1ncn2C	intermediate	5000.0

Table 2: Selected features for preparation of dataset for EDA

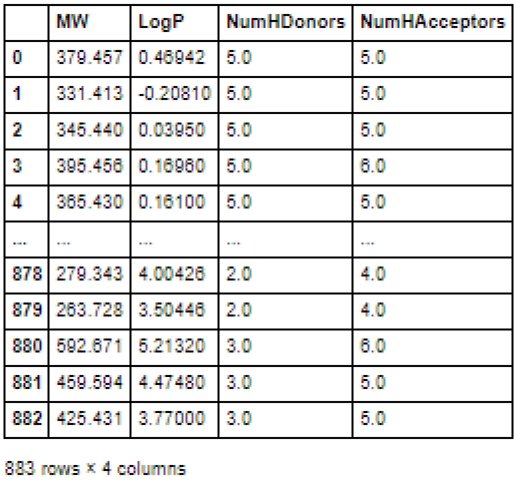
4.2.2.2. **Calculation of Lipinski Descriptors:** Pfizer scientist Christopher Lipinski devised a set of guidelines for assessing the drug-likeness of compounds. The Absorption, Distribution, Metabolism, and Excretion (ADME) profile, often known as the pharmacokinetic profile, is used to determine drug similarity. Lipinski evaluated all orally active FDA-approved medications when he came up with the Rule-of-Five, often known as Lipinski's Rule. According to the rule, any drug that fulfils two or more of the set conditions, it is expected to have poor permeability and absorption [59]. The rule can be able to categorize and remove molecules that are predicted to have poor drug-likeliness [60].



Molecular weight	Greater than 500 Da
Calculated log P values	Above 5
Hydrogen Bond donors	More than 5
Hydrogen Bond Acceptors	More than 10

Table 3: Theshold values for Lipinksi's Descriptors

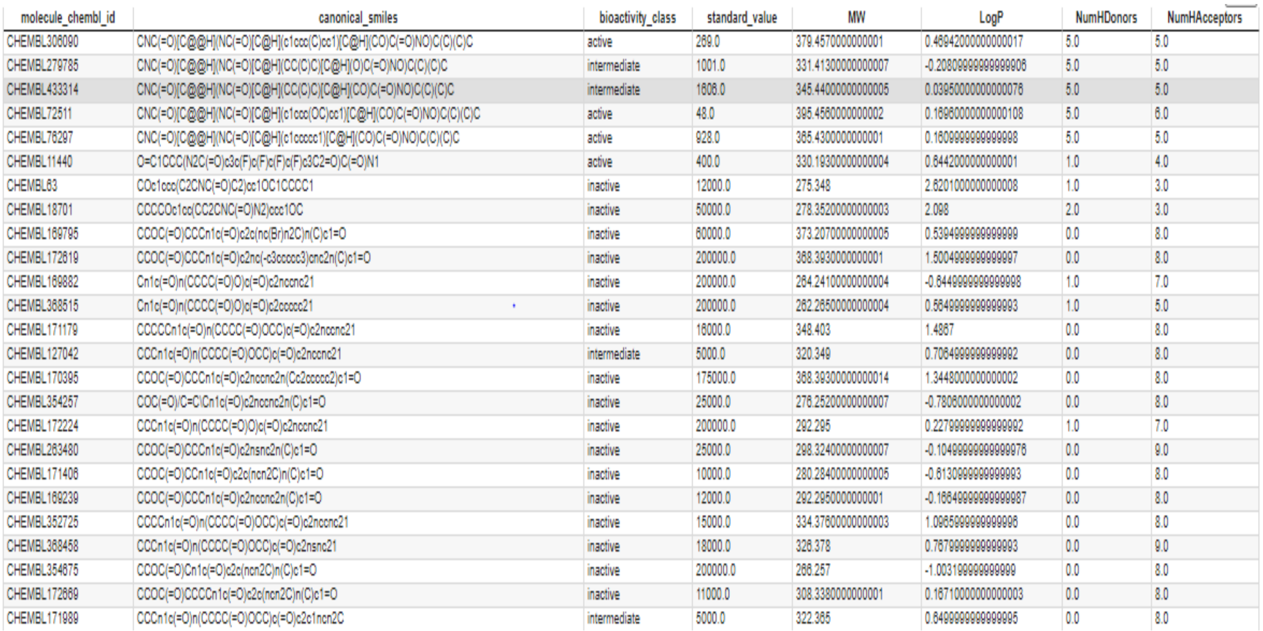
Here the descriptors are used for statistical distribution and analysis of the compounds under study. All for Lipinski Descriptors are calculated and then integrated into the created data subset for the analysis.



	MW	LogP	NumHDonors	NumHAcceptors
0	379.457	0.46942	5.0	5.0
1	331.413	-0.20810	5.0	5.0
2	345.440	0.03950	5.0	5.0
3	395.456	0.16960	5.0	6.0
4	365.430	0.16100	5.0	5.0
...	...	...	...	...
878	279.343	4.00426	2.0	4.0
879	263.728	3.50446	2.0	4.0
880	592.671	5.21320	3.0	6.0
881	459.594	4.47480	3.0	5.0
882	425.431	3.77000	3.0	5.0

883 rows × 4 columns

Fig 7: Values of Lipinski Descriptors for various compounds in the dataset



molecule_chembl_id	canonical_smiles	bioactivity_class	standard_value	MW	LogP	NumHDonors	NumHAcceptors
CHEMBL306090	CNC(=O)[C@@H](NC(=O)[C@H](c1ccc(C)cc1)[C@H](CO)C(=O)NO)C(C)(C)C	active	269.0	379.4570000000001	0.4694200000000017	5.0	5.0
CHEMBL279785	CNC(=O)[C@@H](NC(=O)[C@H](CC(C)C)[C@H](O)C(=O)NO)C(C)(C)C	intermediate	1001.0	331.4130000000007	-0.2080999999999906	5.0	5.0
CHEMBL433314	CNC(=O)[C@@H](NC(=O)[C@H](CC(C)C)[C@H](CO)C(=O)NO)C(C)(C)C	intermediate	1606.0	345.4400000000005	0.0395000000000076	5.0	5.0
CHEMBL72511	CNC(=O)[C@@H](NC(=O)[C@H](c1ccc(OC)cc1)[C@H](CO)C(=O)NO)C(C)(C)C	active	48.0	395.4560000000002	0.1696000000000108	5.0	6.0
CHEMBL76297	CNC(=O)[C@@H](NC(=O)[C@H](c1ccccc1)[C@H](CO)C(=O)NO)C(C)(C)C	active	928.0	365.4300000000001	0.1609999999999998	5.0	5.0
CHEMBL11440	O=C1CCC(N2C(=O)c3c(F)c(F)c(F)c(F)c3C2=O)C(=O)N1	active	400.0	330.1930000000004	0.6442000000000001	1.0	4.0
CHEMBL63	COc1ccc(C2CNC(=O)C2)cc1OC1CCCC1	inactive	12000.0	275.348	2.6201000000000008	1.0	3.0
CHEMBL18701	CCCCOc1cc(CC2CNC(=O)N2)ccc1OC	inactive	50000.0	278.3520000000003	2.098	2.0	3.0
CHEMBL169795	CCOC(=O)CCCn1c(=O)c2c(nc(Br)n2C)n(C)c1=O	inactive	60000.0	373.2070000000005	0.5394999999999999	0.0	8.0
CHEMBL172619	CCOC(=O)CCCn1c(=O)c2nc(-c3ccccc3)cnc2n(C)c1=O	inactive	200000.0	368.3930000000001	1.5004999999999997	0.0	8.0
CHEMBL169882	Cn1c(=O)n(CCCC(=O)O)c(=O)c2nccnc21	inactive	200000.0	264.2410000000004	-0.6449999999999998	1.0	7.0
CHEMBL368515	Cn1c(=O)n(CCCC(=O)O)c(=O)c2ccccc21	inactive	200000.0	262.2650000000004	0.5649999999999993	1.0	5.0
CHEMBL171179	CCCCCn1c(=O)n(CCCC(=O)OCC)c(=O)c2nccnc21	inactive	16000.0	348.403	1.4867	0.0	8.0
CHEMBL127042	CCCn1c(=O)n(CCCC(=O)OCC)c(=O)c2nccnc21	intermediate	5000.0	320.349	0.7064999999999992	0.0	8.0
CHEMBL170395	CCOC(=O)CCCn1c(=O)c2nccnc2n(Cc2ccccc2)c1=O	inactive	175000.0	368.3930000000014	1.3448000000000002	0.0	8.0
CHEMBL354257	COC(=O)/C=C\Cn1c(=O)c2nccnc2n(C)c1=O	inactive	25000.0	276.2520000000007	-0.7806000000000002	0.0	8.0
CHEMBL172224	CCCn1c(=O)n(CCCC(=O)O)c(=O)c2nccnc21	inactive	200000.0	292.295	0.2279999999999992	1.0	7.0
CHEMBL263480	CCOC(=O)CCCn1c(=O)c2nsnc2n(C)c1=O	inactive	25000.0	298.3240000000007	-0.1049999999999976	0.0	9.0
CHEMBL171406	CCOC(=O)CCn1c(=O)c2c(ncn2C)n(C)c1=O	inactive	10000.0	280.2840000000005	-0.6130999999999993	0.0	8.0
CHEMBL169239	CCOC(=O)CCCn1c(=O)c2nccnc2n(C)c1=O	inactive	12000.0	292.2950000000001	-0.1664999999999987	0.0	8.0
CHEMBL352725	CCCCn1c(=O)n(CCCC(=O)OCC)c(=O)c2nccnc21	inactive	15000.0	334.3760000000003	1.0965999999999996	0.0	8.0
CHEMBL368458	CCCn1c(=O)n(CCCC(=O)OCC)c(=O)c2nsnc21	inactive	18000.0	326.378	0.7679999999999993	0.0	9.0
CHEMBL354675	CCOC(=O)Cn1c(=O)c2c(ncn2C)n(C)c1=O	inactive	200000.0	266.257	-1.003199999999999	0.0	8.0
CHEMBL172669	CCOC(=O)CCCCn1c(=O)c2c(ncn2C)n(C)c1=O	inactive	11000.0	308.3380000000001	0.1671000000000003	0.0	8.0
CHEMBL171989	CCCn1c(=O)n(CCCC(=O)OCC)c(=O)c2c1ncn2C	intermediate	5000.0	322.365	0.6499999999999995	0.0	8.0

Fig 8: Final dataset for model building

4.2.2.3. **Conversion IC50 to pIC50:** pIC50 values are the negative logarithmic values of IC50 i.e. -log10(IC50). Calculation of pIC50 values makes the data more uniform for analysis and plotting of graphs. It will urge you to consider over your potency data in logarithmic scales rather than arithmetic scales. It's difficult to report the accuracy of IC50 determinations.

4.2.2.4. **Exploratory Data Analysis:** The data subset containing the descriptor values is then used for some graphical and statistical analysis to achieve a superior understanding of the data points through visualizations. With the use of summary statistics and graphical representations, EDA is a crucial step in doing preliminary investigations on data for uncovering patterns, detecting anomalies, testing hypotheses, & looking for assumptions [61].

4.2.2.5. **Calculation of Fingerprint Descriptors:** The paDEL-Descriptors are numerical values derived from the structural and chemical properties of the compounds.

These numerical values are used as the input for the model building in QSAR modeling and can be used to predict the bioactivity of novel compounds. The descriptor is available in the library format and provides around 43 algorithms for the molecular descriptors and 7 fingerprint algorithms. It is open source software available for free which makes it more accessible for application over other available softwares [62].

In this experiment 881 descriptors of pubChem Fingerprint were taken for analysis and model building.

4.2.3. **Validating the QSAR model:** Validating the model that has been created is an important step in order to assure that the created model would be able to make precise and reliable predictions in the future. Most commonly used parameters to analyze the performance of the model are Pearson's correlation coefficient (r) and root mean squared error (RMSE).

The value of Pearson’s coefficient is the representation of the degree of relationships amongst the features in consideration and can have the values ranging from -1 to +1. The positive value shows a positive correlation and vice versa. The Root Mean Squared Error is a parameter that is used to evaluate the possible error of the model.

Few other methods like F-statistic, standard deviation, Y-scrambling test are additionally used to increase the validation assessment.

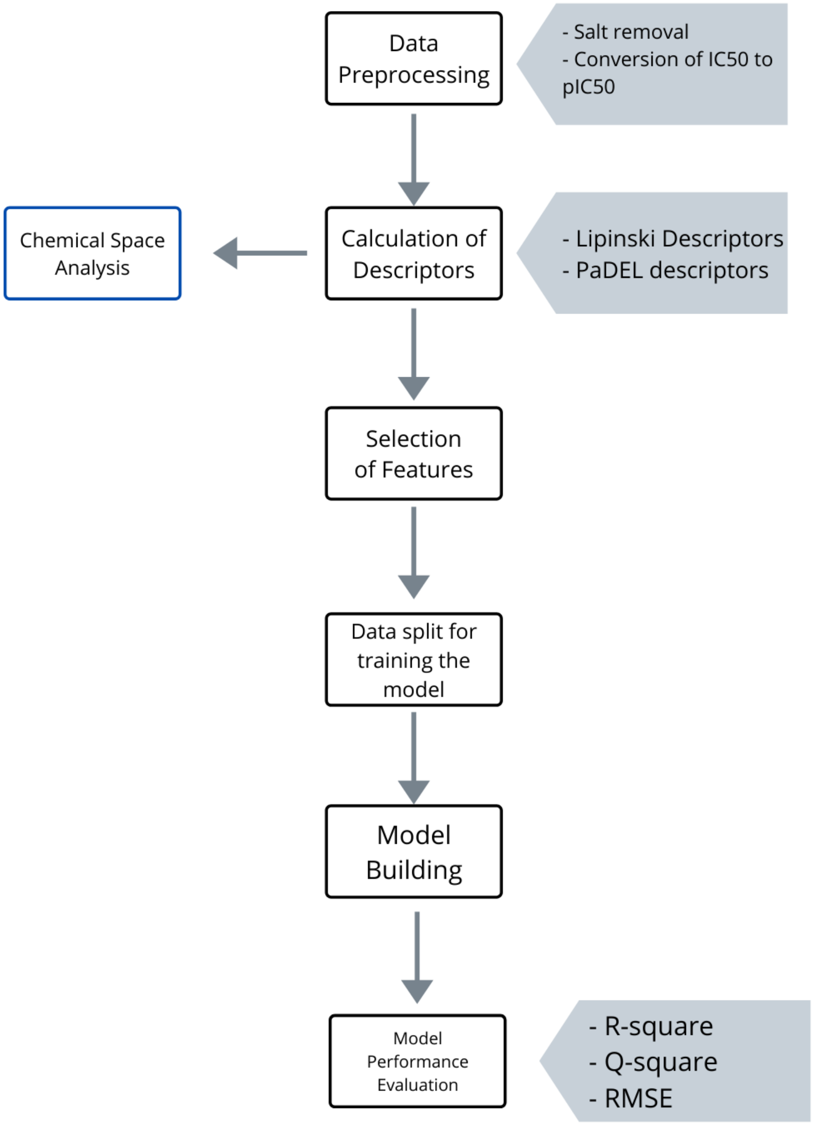


Fig 9: General scheme for model building and validation for QSAR

Chapter 5

Results and Discussion

5.1. Results of Exploratory Data Analysis

It is observed through the scatter plot of MW vs LogP (Fig) that the two bioactivity classes are found in almost similar chemical spaces. The QSAR model involves considering a large number of compounds that can be represented by descriptors. The chemical space analysis show hig correlation between the active and inactive compounds.

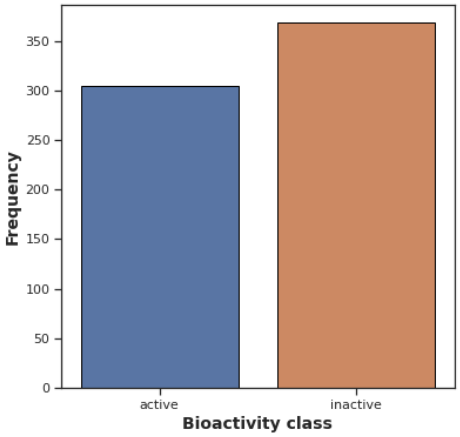


Fig 10: Frequency plot of the 2 bioactivity classes

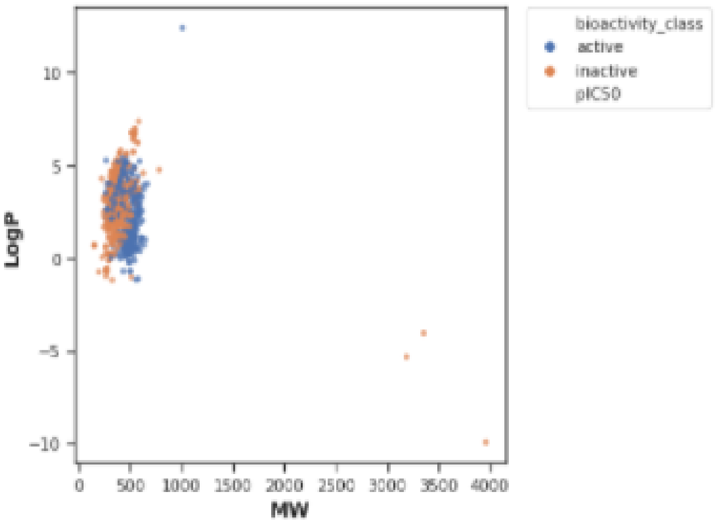


Fig 11: Scatter plot of MW versus LogP values

5.2. Interpretation of Statistical Results

The different box plots were created in order to understand the chemical space of the TNF-alpha inhibitors and get inferences about the SAR by using the Lipinski's Descriptors to establish the relationship. Such a chemical space analysis provides insights about the common characteristics of the compounds under study. EDA was performed on these descriptors to understand the statistical inference of the distribution across the chemical space. Molecular size (depicted by MW) is the common parameter used, since the understanding of the molecular size of the compound is required to predict its ability to pass through the lipid bilayer. It is observed that most of the inhibitors lie within the range of 400-500 Da. LogP is also a commonly used for analyzing the penetration and permeation capabilities of the compounds. It gives an understanding of the liphobilicity of the compounds. The active compounds are spread over the space of 1 to 3 whereas the inactive compounds are spread over 2 to 4. The HB-donors and HB-acceptors are the measures to understand the hydrogen bonding capacity of the compounds

Box Plots

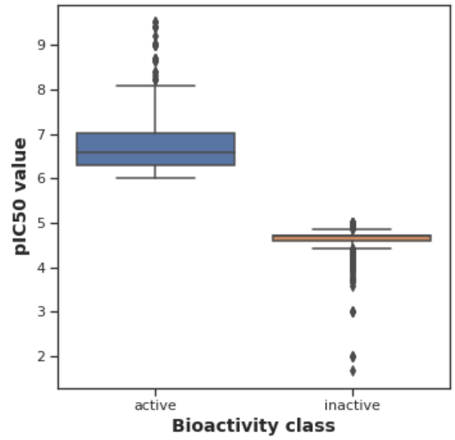


Fig 12: Box plot of pIC50 value for the 2 classes

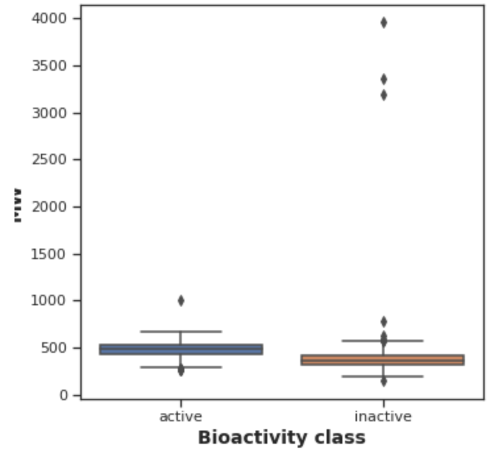


Fig 13: Box plot of MW value for the 2 classes

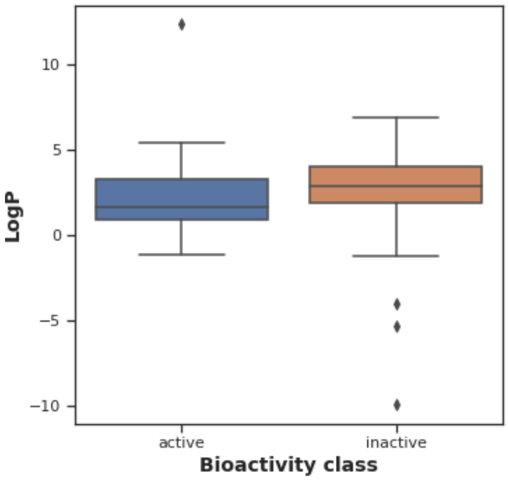


Fig 14: Box plot of LogP value for the 2 classes

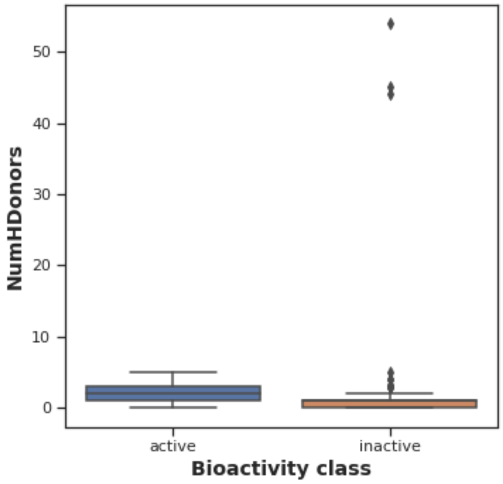


Fig 15: Box plot of NumHDonors value for the 2 classes

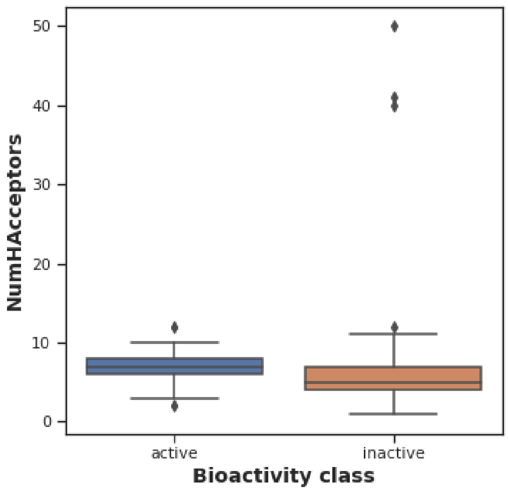


Fig 15: Box plot of NumHAcceptors value for the 2 classes

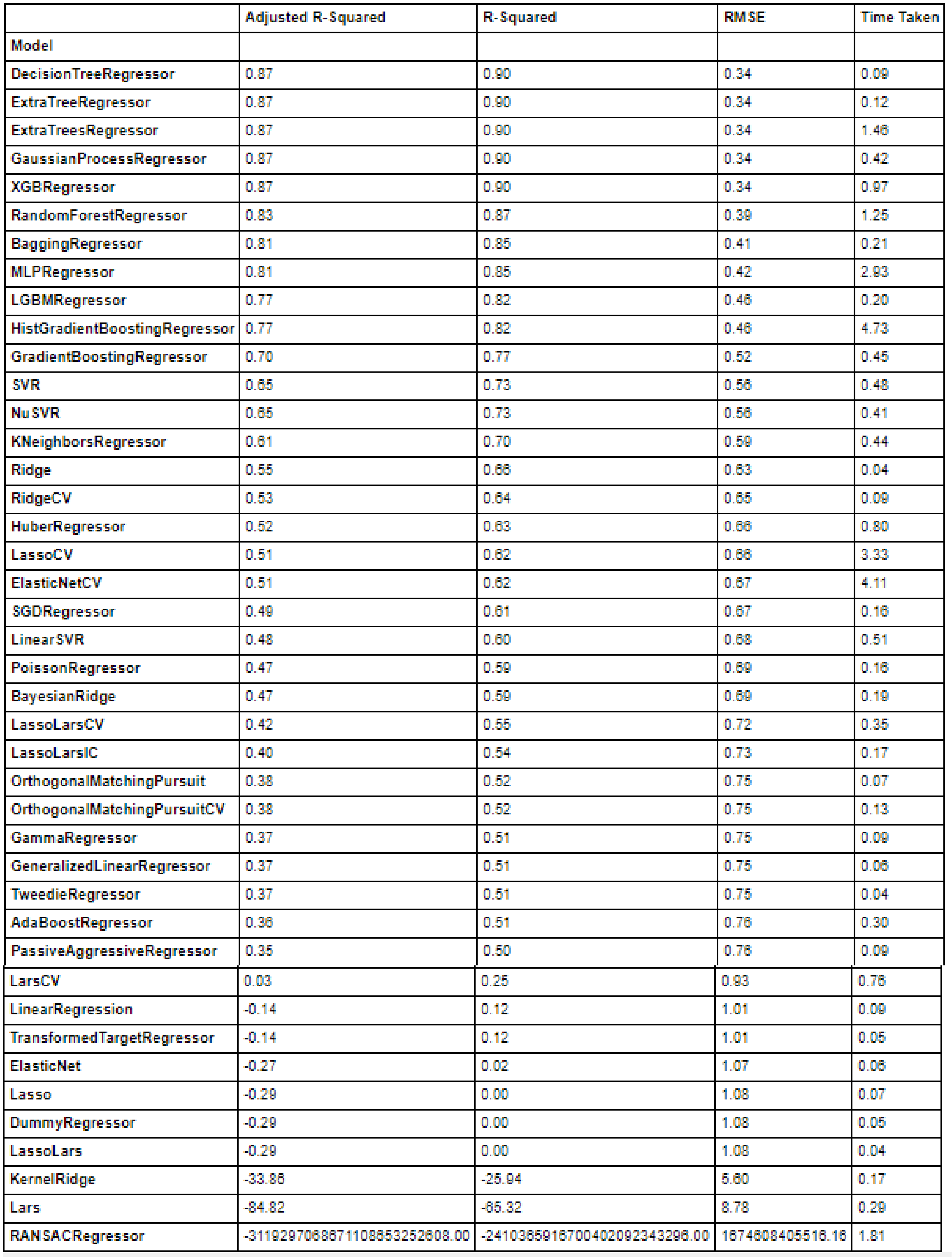
5.3. Discussion

A data set initially consisting of 1690 compounds had been narrowed down to be used to construct the QSAR model. Different fingerprints were used to convert the structural description into numerical values. After the data cleaning and filter selection, the dataset was split into a 80/20 ratio that comprised of 80% training data set and 20% test dataset. Random forest model was used to test the predictability of the dataset. The value of Pearson's Coefficient obtained was **0.591** which indicates a strong model performance.

While performing the drug discovery processes, it is very essential to know about the ADMET properties of the compound(s) of interest. This information can be easily extracted from data retrieved from ChEMBL and extracting Lipinski Descriptors for the compound selected to analyse the bioactivity. The amount of data availability and ML has reduced the time and efforts involved in the examination of a large number of information present in the form of journals or research papers. The properties like molecular weight (MW), water partition coefficient (logP), ability to form hydrogen bonds are essential determinants of the drug-like properties of the compounds. The compounds having similar chemical structures are also likely to have similar bioactivity. However, there might be some fundamental differences that can cause them to have slightly different applications. This can be analyzed by a technique called Multiparameter Optimisation wherein we can analyze a compound’s drug likeness, safety and efficacy against multiple parameters. Optimising biological characteristics while also ensuring that the structure's property profile is drug-like, for example, might be done utilising the Lipinski rule-of-five criterion. Advances in the ML and AL technologies can resolve these concerns provided considerable amount of quality data to approve the effectiveness in the drug discovery process.

QSAR models provide the relationships between the structural and biological characteristics of the compounds. The tried regression model did not show the best results for the dataset, therefore, different regression models were tested for comparisons which revealed that the decision tree regressor could be the best model for the predictions of biological activity Fig.16-18.

Comparing various regression models



	Adjusted R-Squared	R-Squared	RMSE	Time Taken
Model				
DecisionTreeRegressor	0.87	0.90	0.34	0.09
ExtraTreeRegressor	0.87	0.90	0.34	0.12
ExtraTreesRegressor	0.87	0.90	0.34	1.46
GaussianProcessRegressor	0.87	0.90	0.34	0.42
XGBRegressor	0.87	0.90	0.34	0.97
RandomForestRegressor	0.83	0.87	0.39	1.25
BaggingRegressor	0.81	0.85	0.41	0.21
MLPRegressor	0.81	0.85	0.42	2.93
LGBMRegressor	0.77	0.82	0.46	0.20
HistGradientBoostingRegressor	0.77	0.82	0.46	4.73
GradientBoostingRegressor	0.70	0.77	0.52	0.45
SVR	0.65	0.73	0.56	0.48
NuSVR	0.65	0.73	0.56	0.41
KNeighborsRegressor	0.61	0.70	0.59	0.44
Ridge	0.55	0.66	0.63	0.04
RidgeCV	0.53	0.64	0.65	0.09
HuberRegressor	0.52	0.63	0.66	0.80
LassoCV	0.51	0.62	0.66	3.33
ElasticNetCV	0.51	0.62	0.67	4.11
SGDRegressor	0.49	0.61	0.67	0.16
LinearSVR	0.48	0.60	0.68	0.51
PoissonRegressor	0.47	0.59	0.69	0.16
BayesianRidge	0.47	0.59	0.69	0.19
LassoLarsCV	0.42	0.55	0.72	0.35
LassoLarsIC	0.40	0.54	0.73	0.17
OrthogonalMatchingPursuit	0.38	0.52	0.75	0.07
OrthogonalMatchingPursuitCV	0.38	0.52	0.75	0.13
GammaRegressor	0.37	0.51	0.75	0.09
GeneralizedLinearRegressor	0.37	0.51	0.75	0.06
TweedieRegressor	0.37	0.51	0.75	0.04
AdaBoostRegressor	0.36	0.51	0.76	0.30
PassiveAggressiveRegressor	0.35	0.50	0.76	0.09
LarsCV	0.03	0.25	0.93	0.76
LinearRegression	-0.14	0.12	1.01	0.09
TransformedTargetRegressor	-0.14	0.12	1.01	0.05
ElasticNet	-0.27	0.02	1.07	0.06
Lasso	-0.29	0.00	1.08	0.07
DummyRegressor	-0.29	0.00	1.08	0.05
LassoLars	-0.29	0.00	1.08	0.04
KernelRidge	-33.86	-25.94	5.60	0.17
Lars	-84.82	-65.32	8.78	0.29
RANSACRegressor	-3119297068871108653252608.00	-2410365916700402092343296.00	1674608405516.16	1.81

Table 4: Comparison table for various regression models and the validation parameter values

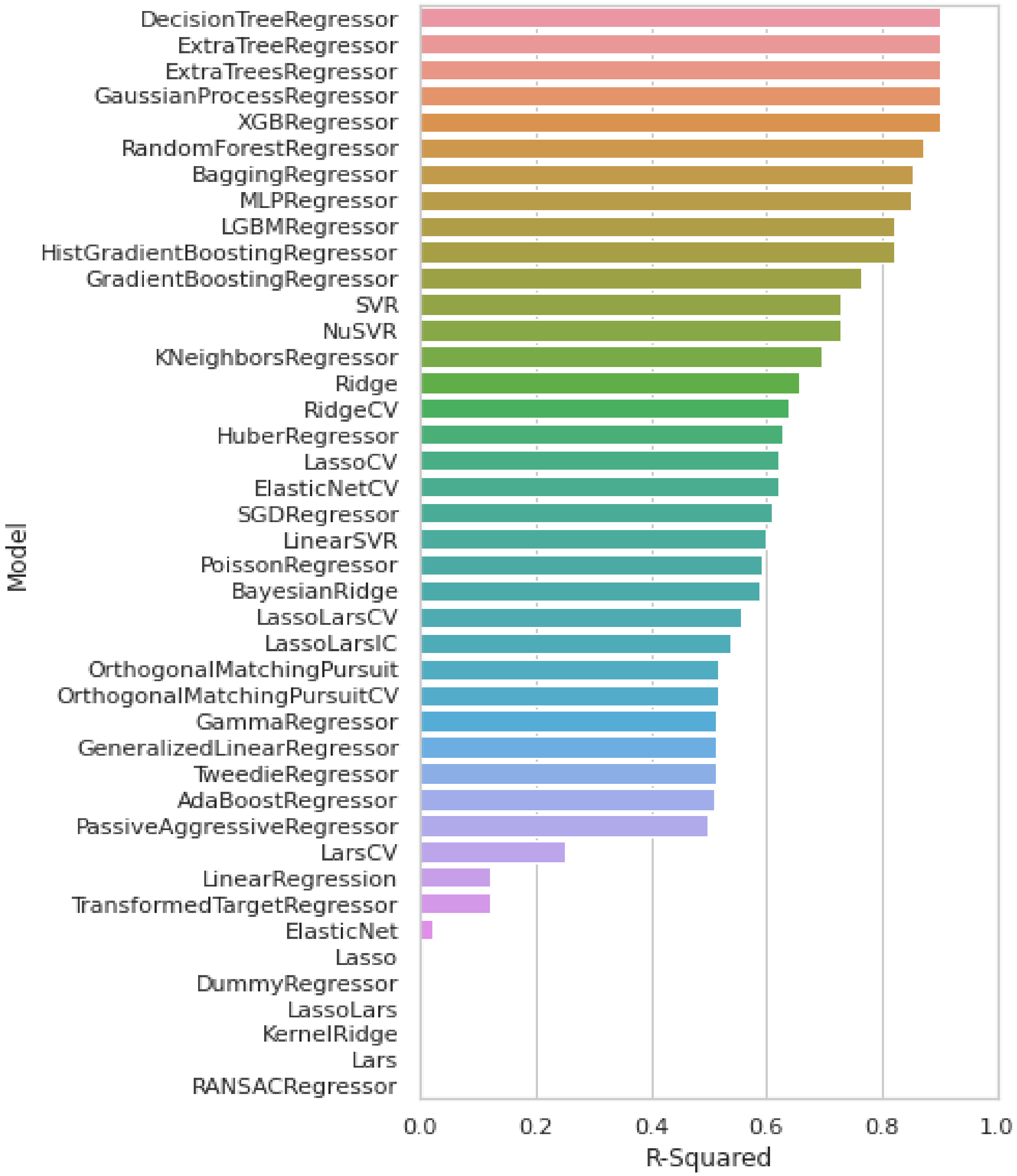


Fig 16: Comparison of various regression models vs their R-square values

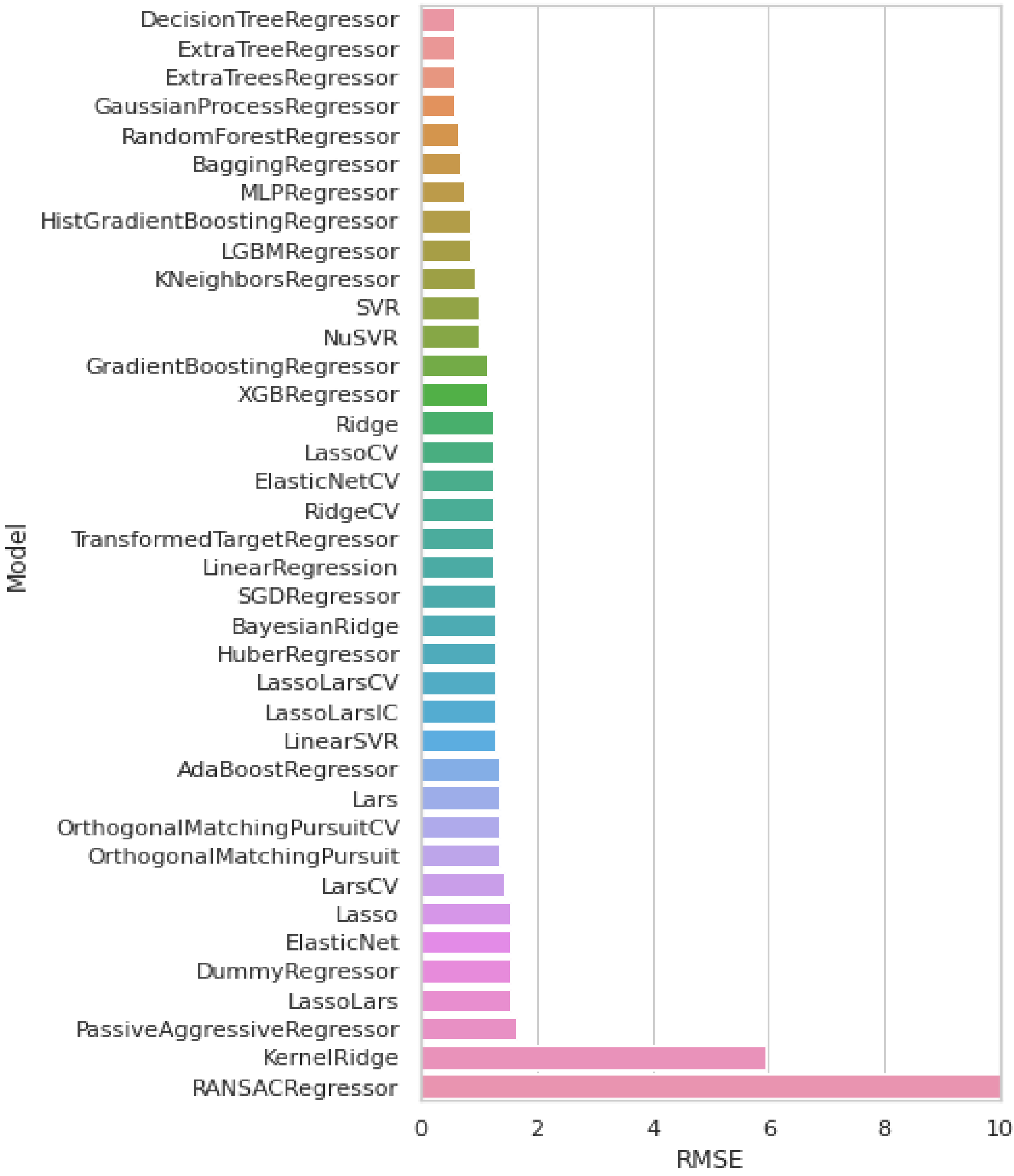


Fig 18: Comparison of various regression models vs their RMSE values

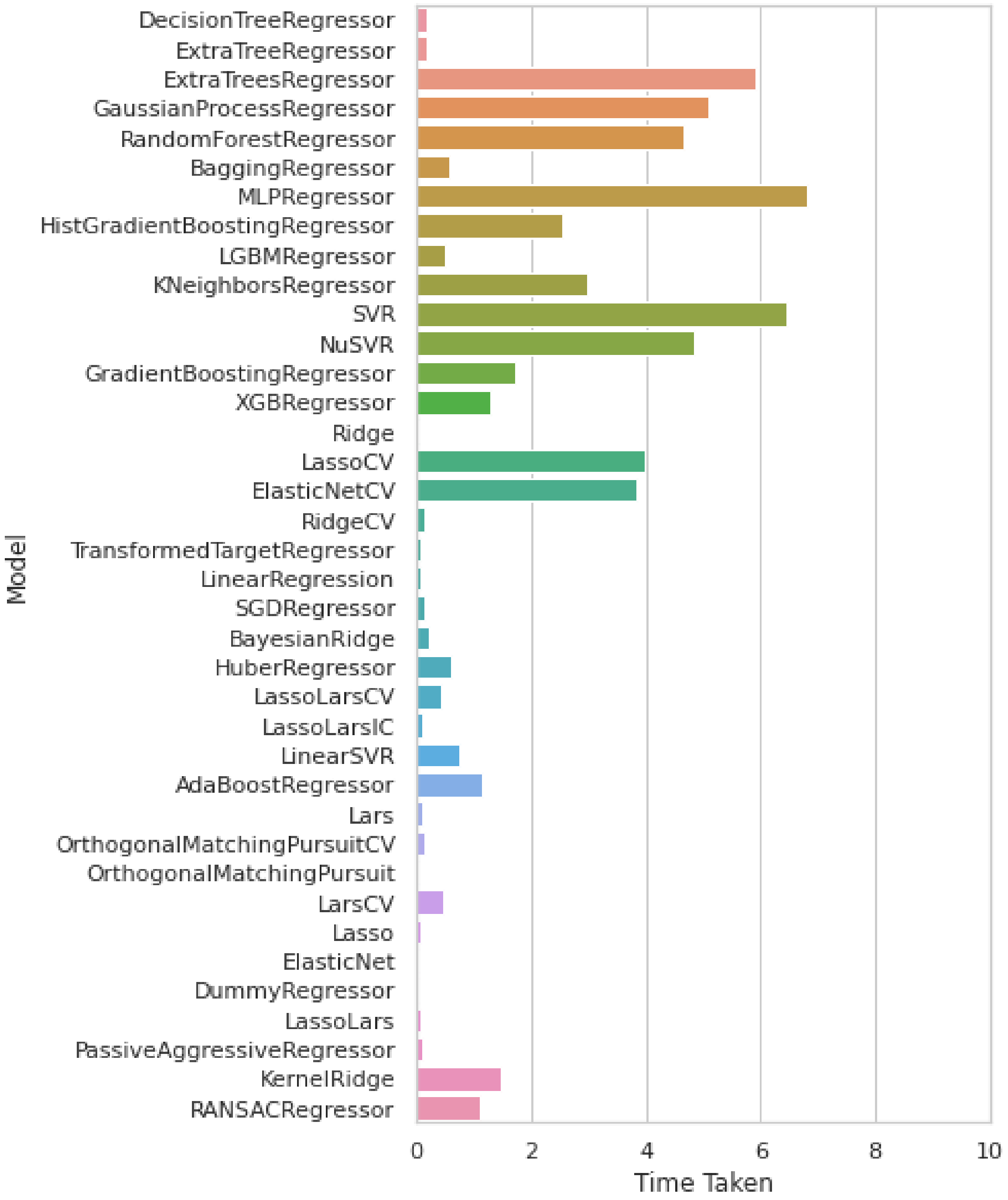


Fig 18: Comparison of various regression models vs their time taken for execution

Conclusion

Traditional databases are only useful in handling and storing small amounts of data. As the data becomes complex and unstructured, traditional databases fail to extract knowledge from it. Big data technology comes with a promise to provide using terabytes of data to derive useful insights to the medical field, thereby leading to enhancement of the clinical outcomes and improvement of potential healthcare outcomes. There is always a concern with the compromise on data quality to manage hard deadlines. Even after proper analysis, data validation becomes the issue. To overcome the preceding challenges, the data needs to be analyzed efficiently and at minimum costs. Machine Learning serves as the best possible solution towards reducing the increasing costs as well as in the establishment of better doctor-physician relationships. ML and BDA can together be used for several healthcare-based applications including treatments for cancer, many rare diseases leading the way towards personalized medications. Efficient data mining techniques can provide with endless possibilities for data model analysis to reveal patterns that can be used by healthcare professionals in forecasting, diagnosis and treatment of patients.

One of the greatest drawbacks relating to big data is the negligence of protecting privacy, majorly for the data obtained from confidential medical records. Even though there are rules protecting the privacy of medical records, many of them are not applicable to the transfer of big data. Several academics and healthcare professionals agree that present privacy policies need to be overhauled in order to safeguard patients while e also allowing analysts to do successful analysis. While the potential to forecast future medical difficulties is seen as a benefit by some, big data also pose the prospect of medical practitioners being replaced. Various experts are concerned that the rise of big data would underestimate the potential of doctors, leading patients seeking solutions from technology rather than a certified physician. However, as technology advances, the downsides must be considered in order to provide a patient and doctor experience that is both efficient and safe.

The medical industry has to undergo many challenges involved during the pre-clinical phases and clinical phases of drug discovery. The presence of the huge datasets available would enhance the effectiveness and reduce the time invested in the early stages of drug development. In addition to this the use of ML would make the process of decision making more efficient and fast.

Future scope of research

The data gathered from various sectors is mainly unstructured and versatile. It needs to be edited along different dimensions within or across organizations to receive the desired outcomes. Different frameworks for big data analysis are still unexplored to fill the gaps in bringing big data and healthcare together. Proper analysis across organizations requires proper storage facilities for the vast amount of data captured. The analyzed data can be further used in the betterment of the field of personalized medications thereby, allowing the timely diagnosis and cure of rare diseases. The efficient use of technology in the right direction might also help in the reduction of costs for treatments. Healthcare organizations need to work on more effective use of predictive data analysis and link data from multiple sources. With increasingly rising sources of data, there is a requirement for more attention on the new ways of preserving privacy and ethical concerns. Companies can use the rapidly increasing data in critical ways to create new services in response to observable patterns.

There is requirement of technocrats with expert knowledge for using advance Ml algorithms and BDA tools to analyze complex data types and convert them into useful predictions. Such, changes in the field of medicine and life science can help in treatments and cures for life threatening diseases. BDA can transform the possibilities of biomedical research and discoveries.

References

[1] Sinha, S., & Vohora, D. (2018). *Drug Discovery and Development. Pharmaceutical Medicine and Translational Clinical Research, 19–32.* doi:10.1016/b978-0-12-802103-3.00002-x

[2] Deore, Amol & Dhumane, Jayprabha & Wagh, Rushikesh & Sonawane, Rushikesh. (2019). The Stages of Drug Discovery and Development Process. Asian Journal of Pharmaceutical Research and Development. 7. 62-67. 10.22270/ajprd.v7i6.616.

[3] Shayne CG. Introduction: drug Discovery in the 21stCentury. Drug Discovery Handbook, Wiley Press, 2005; 1-10.

[4] Smith GC, OíDonnel JT. The Process of New Drug Discovery and Development, Eds., 2nd edition, Informa Healthcare, New York 2006.

[5] Moffat J, Vincent F, Lee J, Eder J, Prunotto M. Opportunities and challenges in phenotypic drug discovery: an industry perspective. Nature Reviews Drug Discovery, 2017; 16(8):531-543.

[6] Greene CS, Tan J, Ung M, Moore JH, Cheng C.: Big Data Bioinformatics. J Cell Physiol. 2014 Dec;229(12):1896-900. doi: 10.1002/jcp.24662. Erratum In: J Cell Physiol. 2016 Jan;231(1):257. PMID: 24799088; PMCID: PMC5604462.

[7] Brown, N., Cambruzzi, J., Cox, P. J., Davies, M., Dunbar, J., Plumbley, D., … Sheppard, D. W. (2018). *Big Data in Drug Discovery. Progress in Medicinal Chemistry, 277–356.* doi:10.1016/bs.pmch.2017.12.003

[8] Sinha, S., & Vohora, D. (2018). *Drug Discovery and Development. Pharmaceutical Medicine and Translational Clinical Research, 19–32.* doi:10.1016/b978-0-12-802103-3.00002-x

[9] Pina, A. S., Hussain, A., & Roque, A. C. A. (2009). *An Historical Overview of Drug Discovery. Methods in Molecular Biology, 3–12.* doi:10.1007/978-1-60761-244-5_1

[10] Patil, H.K., Seshadri, R.: Big Data Security and Privacy Issues in Healthcare. *2014 IEEE International Congress on Big Data*, pp. 762-765, Anchorage, AK, USA, (2014)

[11] Ye, S.Q. (ed.): Big Data Analysis for Bioinformatics and Biomedical Discoveries (1st Ed.). Chapman and Hall/CRC. (2016)

[12] Villars, R.L. Olofson, C.W. Eastwood, M.: Big Data: What It Is and Why You Should Care, White Paper, IDC, MA, USA (2011)

[13] Scarpati, J.: Big Data Analysis: Storage, Network and Server Challenges. Search Cloud Provider, (2012)

[14] Davenport, Thomas H., Paul B, and Randy B.: How 'Big Data' Is Different. MIT Sloan Management Review 54, no. 1 (Fall 2012)

[15] Sonnati, R.: Improving Healthcare Using Big Data Analytics. International Journal Of Scientific & Technology Research Volume 6, Issue 03, ISSN 2277-8616 (March 2017)

[16] Senthilkumar, S.A., Bharatendara, K.R., Amruta, A.M., Angappa, G., Chandrakumarmangalam S.: Big Data in Healthcare Management: A Review of Literature, American Journal of Theoretical and Applied Business. Vol. 4, No. 2, 2018, pp. 57-69. (2018)

[17] Al-Salim, A. M., Lawey, A.Q., El-Gorashi, T. E. H., Elmirghani, J. M. H.: Energy Efficient Big Data Networks: Impact of Volume and Variety. In: IEEE Transactions on Network and Service Management, vol. 15, no. 1, pp. 458-474 (March 2018)

[18] Pramanik, P.K.D. Saurabh, P., Mukherjee, M.: Healthcare Big Data: A Comprehensive Overview. 10.4018/978-1-5225-7071-4.ch004 (2018)

[19] Al-Salim, A. M., Lawey, A.Q., El-Gorashi, T. E. H., Elmirghani, J. M. H.: Greening Big Data Networks: Velocity Impact. In: ET Digital Library, Vol 12, Issue 3, pp. 126 – 135 (June 2018)

[20] Abawajy, J.: Comprehensive Analysis of Big Data Variety Landscape. International Journal of Parallel Emergent and Distributed Systems. Vol: 30 (2014)

[21] Bresnick, J.: Top 10 Challenges of Big Data Analytics in Healthcare. (2017)

[22] Kulkarni, A. J., Siarry, P., Singh, P. K., Abraham, A., Zhang, M., Zomaya, A., & Baki, F. (eds.): Big Data Analytics in Healthcare. Studies in Big Data. (2020)

[23] Raghupathi, W., Raghupathi, V.: Big data Analytics in Healthcare: Promise and potential. Health Information Science and Systems. 2. 3. 10.1186/2047-2501-2-3 (2014).

[24] Dash, S., Shakyawar, S., Sharma, M., Kaushik, S.: Big Data in Healthcare: Management, Analysis and Future Prospects. Journal of Big Data. 6. 10.1186/s40537-019-0217-0 (2019)

[25] Costa, F.F.: Big data in genomics: challenges and solutions. G.I.T. Lab. J. 11– 12, 1–4 (2012).

[26] Costa, F. F.: Big data in Biomedicine. Drug Discovery Today, 19(4), 433–440. (2014).

[27] Archenaa, J., Anita, E. A. M.: A Survey of Big Data Analytics in Healthcare and Government, Procedia Comput. Sci. 50, 408–413 (2015).

[28] El aboudi, N., & Benhlima, L.: Big Data Management for Healthcare Systems: Architecture, Requirements, and Implementation. Advances in Bioinformatics, 2018, 1–10. (2018)

[29] Ngiam, K. Y., & Khor, I. W.: Big Data and Machine Learning Algorithms for Health-Care Delivery. The Lancet Oncology, 20(5), e262–e273 (2019).

[30] Ţăranu, I.: Data Mining in Healthcare: Decision Making and Precision.” Database Systems Journal 6, 33-40 (2016)

[31] Belle, A., Thiagarajan, R., Soroushmehr, S. M., Navidi, F., Beard, D. A., & Najarian, K.: Big Data Analytics in Healthcare. BioMed research international, 2015, 370194. (2015).

[32] Dhar V.: Big Data and Predictive Analytics in Health Care. Big Data. Sep;2(3):113-6. (2014)

[33] Lepenioti, K., Bousdekis, A., Apostolou, D., Mentzas, G.: Prescriptive Analytics: A Survey of Approaches and Methods: BIS 2018 International Workshops, Berlin, Germany, July 18–20, 2018, Revised Papers. (2019).

[34] Lopes, J., Guimarães, T., & Santos, M. F.: Predictive and Prescriptive Analytics in Healthcare: A Survey. Procedia Computer Science, 170, 1029–1034. (2020).

[35] Ko, I., Chang, H,:Interactive Visualization of Healthcare Data Using Tableau. Healthcare Informatics Research. 23. 349. 10.4258/hir.2017.23.4.349.(2017)

[36] Alanazi, H. O., Abdullah, A. H., & Qureshi, K. N.: A Critical Review for Developing Accurate and Dynamic Predictive Models Using Machine Learning Methods in Medicine and Health Care. Journal of Medical Systems, 41(4) (2017)

[37] Bhardwaj, R., Nambiar, A. R., & Dutta, D.: A Study of Machine Learning in Healthcare. 2017 IEEE 41st Annual Computer Software and Applications Conference (COMPSAC). (2017).

[38] S.B. Kotsiantis, “Supervised Machine Learning: A Review of Classification Techniques”, Informatica 31 (2007) 249-268

[39] Podgorelec, V., Kokol, P., Stiglic, B., & Rozman, I. (2002). *Journal of Medical Systems, 26(5), 445–463.* doi:10.1023/a:1016409317640

[40] D. Lowd, P. Domingos, “Naïve Bayes Models for Probability Estimation”

[41] Vembandasamy, K., Sasipriya, R. and Deepa, E. (2015) Heart Diseases Detection Using Naive Bayes Algorithm. IJISET-International Journal of Innovative Science, Engineering & Technology, 2, 441-444.

[42] Sureskumar, Kalaiselvi. (2017). Naive Bayesian Classification Approach in Health care Applications.

[43] S. V. M. Vishwanathan and M. Narasimha Murty, "SSVM: a simple SVM algorithm," *Proceedings of the 2002 International Joint Conference on Neural Networks. IJCNN'02 (Cat. No.02CH37290),* 2002, pp. 2393-2398 vol.3, doi: 10.1109/IJCNN.2002.1007516.

[44] X. Zhu, A. B. Goldberg, “*Introduction to Semi – Supervised Learning*”, Synthesis Lectures on Artificial Intelligence and Machine Learning, 2009, Vol. 3, No. 1, Pages 1-130

[45] Gu, L., Zhang, X., You, S., Zhao, S., Liu, Z., & Harada, T. (2020). *Semi-Supervised Learning in Medical Images Through Graph-Embedded Random Forest. Frontiers in Neuroinformatics, 14.* doi:10.3389/fninf.2020.601829

[46] R. S. Sutton, “*Introduction: The Challenge of Reinforcement Learning*”, Machine Learning, 8, Page 225-227, Kluwer Academic Publishers, Boston, 1992

[47] L. P. Kaelbing, M. L. Littman, A. W. Moore, “*Reinforcement Learning: A Survey*”, Journal of Artificial Intelligence Research, 4, Page 237-285, 1996

[48] Jee, K., Kim, G.-H.: Potentiality of Big Data in the Medical Sector: Focus on How to Reshape the Healthcare System. Healthcare Informatics Research, 19(2), 79. (2013).

[49] Luo, J., Wu, M., Gopukumar, D., Zhao, Y.: Big Data Application in Biomedical Research and Health Care: A Literature Review. Biomedical Informatics Insights, 8, BII.S31559. (2016).

[50] Varnek, A. and Baskin, I. (2012) Machine learning methods for property prediction in chemoinformatics: Quo Vadis? J. Chem. Inf. Model. 52, 1413–1437

[51] Ali, S.M. et al. (1997) Butitaxel analogues: synthesis and structure-activity relationships. J. Med. Chem. 40, 236–241

[52] Consonni, V., & Todeschini, R. (2009). *Molecular Descriptors. Recent Advances in QSAR Studies, 29–102.* doi:10.1007/978-1-4020-9783-6_3

[53] Kubinyi, H. (1988) Free Wilson analysis. Theory, applications and its relationship to Hansch analysis. Quant. Struct. Act. Relat. 7, 121–133

[54] Gudjonsson, J. E., & Elder, J. T. (2007). *Psoriasis: epidemiology. Clinics in Dermatology, 25(6), 535–546.* doi:10.1016/j.clindermatol.2007.08.007

[55] Jakobsen, M., Stenderup, K., Rosada, C., Moldt, B., Kamp, S., Dam, T. N., … Mikkelsen, J. G. (2009). *Amelioration of Psoriasis by Anti-TNF-α RNAi in the Xenograft Transplantation Model. Molecular Therapy, 17(10), 1743–1753.* doi:10.1038/mt.2009.141

[56] Mehta, N. N., Teague, H. L., Swindell, W. R., Baumer, Y., Ward, N. L., Xing, X., … Gudjonsson, J. E. (2017). *IFN-γ and TNF-α synergism may provide a link between psoriasis and inflammatory atherogenesis. Scientific Reports, 7(1).* doi:10.1038/s41598-017-14365-1

[57] Campa, M., Ryan, C., & Menter, A. (2015). *An overview of developing TNF-α targeted therapy for the treatment of psoriasis. Expert Opinion on Investigational Drugs, 24(10), 1343–1354.* doi:10.1517/13543784.2015.1076793

[58] Anna Gaulton, Louisa J. Bellis, A. Patricia Bento, Jon Chambers, Mark Davies, Anne Hersey, Yvonne Light, Shaun McGlinchey, David Michalovich, Bissan Al-Lazikani, John P. Overington, ChEMBL: a large-scale bioactivity database for drug discovery, *Nucleic Acids Research*, Volume 40, Issue D1, 1 January 2012, Pages D1100–D1107

[59] Turner, J. V., & Agatonovic-Kustrin, S. (2007). *In Silico Prediction of Oral Bioavailability. Comprehensive Medicinal Chemistry II, 699–724.* doi:10.1016/b0-08-045044-x/00147-4

[60] Kenakin, T. P. (2017). *Pharmacology in Drug Discovery. Pharmacology in Drug Discovery and Development, 275–299.* doi:10.1016/b978-0-12-803752-2.00011-9

[61] Behrens, J. T. (1997). Principles and procedures of exploratory data analysis. *Psychological Methods, 2*(2), 131–160. https://doi.org/10.1037/1082-989X.2.2.131

[62] Yap, C. W. (2010). *PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints. Journal of Computational Chemistry, 32(7), 1466–1474.* doi:10.1002/jcc.21707