

**Explaining Developmental Neurotoxicity By
XAI**

A DISSERTATION

SUBMITTED IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE AWARD OF THE DEGREE

OF

Master of Science

In

Biotechnology

Submitted by:

Nakul Tanwar

2K20/MSCBIO/40

Under the supervision of

Prof. Yasha Hasija

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 Nakul Tanwar, Roll Number: 2K20/MSCBIO/40, student of M.Sc. Biotechnology, hereby declare that the work which is presented in the Major Project entitled **—Explaining developmental neurotoxicity by XAI** 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: Explainable AI; Are we there yet?

Author Names: Tanwar, Nakul, and Hasija, Yasha

Name of Conference: 2022 IEEE Delhi Section Conference (DELCON)

Conference Date and Venue: 11-13 Feb 2022 at Netaji Subhas University of Technology

Registration: Done

Status of Paper: Published

Date of Paper Communication: 31th December 2021

Date of Paper Acceptance: 16th January 2022

Date of Paper Publication: 21st April 2022

Date: Nakul Tanwar

DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College Of Engineering)
Bhawana Road, Delhi-110042

Certificate

I hereby certify that the Project Dissertation titled “**Explaining Developmental Neurotoxicity By XAI**” which is submitted by **Nakul Tanwar (2K20/MSCBIO/40)**, Department of Biotechnology, Delhi Technological University, Delhi in partial fulfillment of the requirement for the award of the degree of Master of Science is recorded 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 any diploma to this university or elsewhere.

Place: Delhi

Date:

Prof. Yasha Hasija
(Supervisor)
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 to 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 been 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 she has offered me. Her insightful feedback pushed me to sharpen my thinking and brought my work to a higher level.

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

I would like to thank MR. Rajkumar Chakraborty without whom I would not have been able to complete this research, and without whom I would not have made it through my master's degree. I would like to thank the following people for helping with this research project: Miss. Neha Kumari, Mrs. Jaishree Meena, and Mrs. Priya

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

Nakul Tanwar

ABSTRACT:

Chemical exposure can cause formative neurotoxicity, which requires fast and exact testing techniques. Human physiology presents various challenges for current techniques such as human essential cell culture examines, in vivo animal examinations, and tests of animal essential cell cultures. Research in this study used joining explainable artificial intelligence (XAI) with XGBoost AI (ML) models that were prepared to utilize binary classification as a strong mix of datasets to identify genes that may be associated with neurotoxicity. Significant genes were found and connected to the progression of neurotoxicity after SHAP values were effectively integrated into the ML models.

Contents

Candidate's Declaration	2
Certificate	3
Acknowledgement	5
Abstract	6
Contents	7
List of Figures	9
List of Tables	10
CHAPTER 1 INTRODUCTION	10
CHAPTER 2 XAI	11
2.1. XAI Characteristics	11
2.1.1. Clarification	11
2.1.2. Significant	11
2.1.3. Accuracy	12
2.1.4. Information Sufficiency	12
2.2. AI's HIERARCHY	12
2.2.1. Inaugural Wave: Aboriginal Knowledge	12
2.2.2. The subsequent Wave: Quantifiable Education	14
2.2.3. The Ternary Wave: The Transformative Process	15
2.3. XAI's FRAMEWORK	16
2.3.1 SHAP	18
2.3.2 LIME	18
CHAPTER 3 Computation OF THE BLACK-BOX	20

CHAPTER 4 Challenges EXPERIENCED BY XAI	23
CHAPTER 5 APPLICATIONS	26
5.1.1 Monetary Forecasting	26
5.1.2 Medical care	26
5.1.3 Industry	26
5.1.4 Infomation Quality	27
CHAPTER 6	29
6.1. LiteratureOverview	29
6.2. Methodology	30
6.2.1. DatasetExtraction	30
6.2.1.1. Expression_2D	31
6.2.1.2. Expression_3D	
6.2.2. FEEDING DATASET ON MACHINE LEARNING	31
6.2.2.1. APPLYING XAI ON TRAINED ML	32
6.2.2.2. ConfusionMatrix	32
CHAPTER 7 RESULT AND VALIDATION	33
CHAPTER 8 BIOLOGICAL SIGNIFICANCE OF IDENTIFIED GENES	40
CONCLUSION	44
REFERENCES	45

LIST OF FIGURES

Fig 1: Three waves of AI.	13
Fig 2: Delineation of the conversely corresponding connection among intricacy and clarification.	17
Fig 3: The SHAP and LIME systems are in operation	19
Fig 4: Ordinary AI versus XAI.	21
Fig 5: Convergence of XAI with various fields.	22
Fig 6: The quantity of True positive cases is shown by the grey squares in the grid (TP), False-positive (FP) and False negative (FN) models are addressed by dark squares in the grid, while true negative (TN) events are addressed by white squares in a similar way.	27
Fig 7: The plot of the SHAP synopsis shows the main qualities and their effect. The y-axis is utilized to rank the significance of every quality element on the y-hub. Shown on the x-axis are the consequences of a quality's impact on the model's result: expanded or decreased forecast. Red demonstrates a quality's impact is genuinely critical, while blue shows no impact by any means.	33
Fig 8: From the given bar plot, genes of most noteworthy significance were driven by carrying out SHAP values on the prepared models, showing ID4, WWC1, SRXN1, BMP7, GRIN3A, NCF2, DDN, F13A1, RNF207, RNASE1, SRPX, KDSR, CELF4, HIST1H3H, MMP12, SRSF6, PRKCSH, TAP1, DALRD3.	35
Fig 9: Confusion matrix (A) shows the accuracy of the model containing all genes, whereas the confusion matrix (B) shows improved accuracy due to previously identified key genes.	37
Fig 11. (A) shows the accuracy of 86% in which the model selected the key genes, whereas (B) shows the high accuracy of 91.67% when the key genes are again fed to the model as input.	39
Fig 12. SHAP Values of the previous model and validation model	40

List of Tables

Table 1: Developments in each wave.	16
-------------------------------------	----

CHAPTER 1

Introduction:

AI and profound learning are two of the most progressive calculations accessible today. They produce amazing outcomes, yet in addition dispense with the requirement for people to deal with, store, and cycle information [1]. These models' essential goal is to further develop accuracy, speed up processes, and perform dynamic undertakings, all of which increment human dependence on these models. Nonetheless, these models need straightforwardness, objectivity, and clarification, delivering them irrelevant to genuine issues [2]. XAI incorporates a goal of giving the interior usefulness of the layers that adds to the dependability and lucidity of these models. XAI has planned so that it gives the slightest bit of real proof on the side of the result and features include that might impact the ultimate conclusion. As this device is moderately new, its expected applications in an assortment of areas, including estimating, medical services, and industry, should be investigated.

The term 'logical' or 'reasonableness' alludes to a model's capacity of defending its result. Furthermore, reasonableness guarantees that the model works inside the standard and guarantees the exactness of models [3]. Furthermore, it can allude to the total and exact portrayal of the result created by a model. Reasonableness can be named either local (for an unaccompanied example of choices) or global (to fathoming the replica's choice expectation calculation). The developing significance of logic accentuates the basic idea of instruments that help people in grasping the way of behaving of black-box models. These apparatuses often make reference to (XAI) [4].

CHAPTER 2

XAI

The reception of cutting-edge innovations brings about the age of colossal measures of information. Simulated intelligence and its advances (AI and Deep learning) oversee and deal with this information, expanding people's dependence on such models [5]. Because of this reliance, people are in a remarkable situation to discover the exact aim at the back of any result. XAI's role is to ensure that the end client comprehends the thinking at the back of the result, which fortifies the work on the models' unwavering quality.

The XAI is embodied in the accompanying standards:

1. Clarification

The essential target of XAI is to give far-reaching substantiation to support the result. In AI, there are five kinds of explanations: for clients, to construct public trust, and to meet authoritative and consistent necessities [6].

2. Significant

These guidelines suggest that XAI must be valuable for the administrator or organization as far as appreciating a clarification in an assortment of ways and managing genuine outcomes related to different levels of examination [6].

3. Accuracy

Accuracy is basic for XAI; it makes sense of the justification behind the result as well as gives a moment of knowledge into the dynamic interaction for a genuine issue. This will likewise build the end-trust client's in such models [6].

4. Information Sufficiency

Every innovation has its inhibitors; by following this rule, XAI ought to utilize information requirements on preparing information to assist with keeping away from the development of inadmissible bits of knowledge [6].

III. AI'S HIERARCHY

Simulated intelligence that can act and respond because of advances in calculations performs difficult undertakings like characterization, acknowledgment, proposal, understanding, and setup [7]. Computer-based intelligence's execution in the medical services/biomedical area further develops judgment, yet in addition, grows therapy choices. In any case, the maximum capacity of such advancements should be understood, which raises worries about the present estimation of AI [8]. Considering this, the guard progressed research project office (DARPA) fostered the 'three rushes of AI', which gives a system for assessing AI and sums up its benefits and impediments [9].

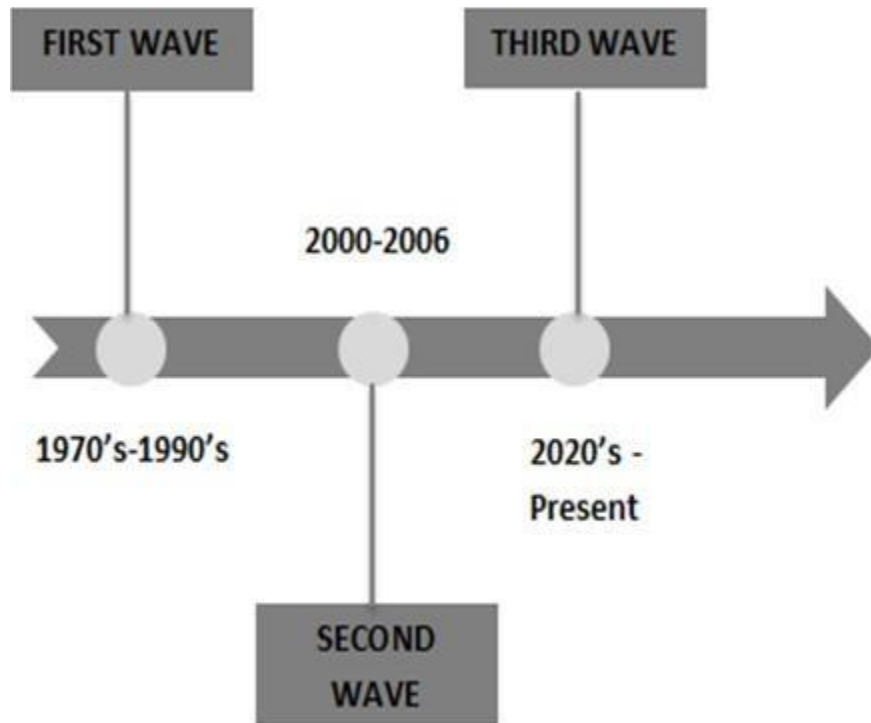


Fig. 1. Three waves of AI.

1. Inaugural Wave: Aboriginal Knowledge

The underlying flood of AI is otherwise called the 'master's framework, as it comprises completely calculations and programming created by human specialists exclusively founded on the information they have. Their essential objective is to furnish these projects with decryptable coherent principles. The main rush of AI depends on coherent standards that require the framework to investigate the most basic boundaries and arrive at a resolution that gives the most suitable activity to take care of an issue. These boundaries are predefined by human specialists, delivering these frameworks unequipped for interpreting novel circumstances [10].

2. The subsequent Wave: Quantifiable Education

The subsequent wave framework requires the most un-human intercession. Even though people make factual models, these models are capable of self-training to deliver high-exactness results [11]. These models win over the general climate and perform a prescient examination that beats human intercession in an assortment of areas. These models are foundations of counterfeit brain networks that are made out of numerous layers that self-train on the preparation of information. The information is input, and each layer processes it alongside the remainder of the organization, delivering precise outcomes as result [12].

To sum up, the second wave framework creates improved results than the primary wave framework. Intruding on these models in an assortment of fields brings about glorious results. These results do not just invigorate further developed direction and take out the human mistakes, yet additionally save huge measures of time, cash, and human exertion. Notwithstanding, these models are 'information subordinate,' as the information decides the model's presentation; an unfortunate preparation dataset will disable the model's capacity to learn and adjust to its current circumstance. Besides that, these models miss the mark as far as their capacity to impart the interior utility of the layer to the end client. [13].

3. The Ternary Wave: The Transformative Process

Third-wave AI frameworks create replicas that will explain programs by uncovering consistent standards that administer their direction. Third-wave structures depend on an assortment of particular real models to acquire enhanced comprehension of a circumstance and to make sense of the dynamic cycle. Inward layer usefulness can be made sense of for the client, in this manner settling the issue of straightforwardness in AI and profound learning [14].

At long last, this wave will help AI by giving pellucidity to the administrator, It adds to the model's appeal as dependable. Nonetheless, this rush is very touchy about information standards, which influences model precision as well as requires an adequate measure of information for preparation.

Table I. Developments in each wave.

Waves	Developments
First brandish	AI
Second brandish	DL & ML
Third-brandish	XAI or interpretable DL & ML

Moreover, there is a futuristic wave that is ordinarily self-overseeing. This wave depends on extremely cutting-edge speculation wherein machines will amalgamate all that and foster the capacity to see and fight back because of their current circumstance [15].

IV. XAI'S FRAMEWORK

XAI is another line of man-made consciousness and AI items. The headway of these advancements in an assortment of areas requires human confidence in the models' choices. The models are exceptionally numerical and hard to decode, which lessens their logic and builds their intricacy.

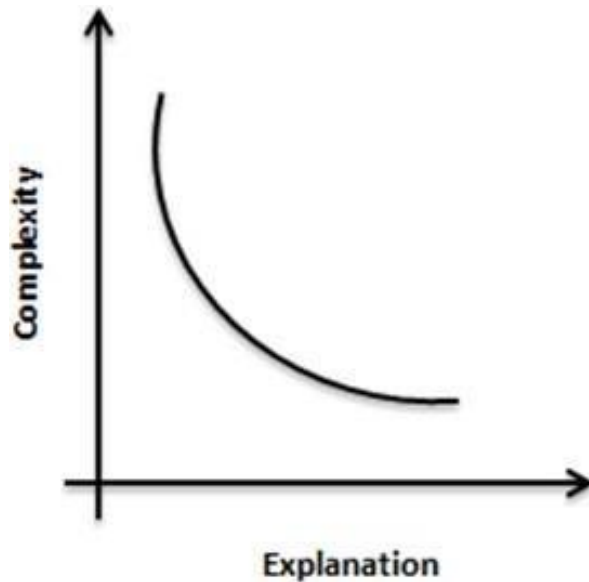


Figure 2. Delineation of the conversely corresponding connection among intricacy and clarification.

The most ideal way to achieve conviction is for the AI model in all honesty. XAI structures are devices that produce examine how a model performs and attempt to make sense of its action. There are different XAI frameworks available, including shapely Additive explanation, Interpretable Model-realist Explanation, ELI5, What-if-device, AIX360, and Skaters, the most frequently used of which are SHAP and LIME. ELI5 is a Python group that forgoes bumbles from AI classifiers. Envision a situation where the device is a Google-made instrument that directs the perception of how AI models work. Open-source includes AIX360 and Skaters which are expansions that enable interpretation of a model, in like manner helping the headway of an interpretable AI structure, which is a large part of the time expected for genuine use cases [16].

1. SHAP

SHAP may be utilized to decipher a wide scope of kinds of models, including vital backslide, tree-based models, AI estimations, and more convoluted models, as significant learning. SHAP relies upon the Shapley regard from the game theory and depicts results that can be changed utilizing various components. It moreover interfaces credit segments and neighborhood improvement utilizing Shapley's worth and its connected developments [17].

2. LIME

LIME is like SHAP, yet speedier concerning evaluation and prepared for figuring out black-box issues with more than one class. LIME's outcome is an ill-defined portrayal of the explanation, containing the components' obligation to the doubt for a nice test. It expects that the classifier fills an undertaking that takes in rough text and makes likelihood values for each class. It can give close help to black-box models by making locally basic data and focusing in on how models advance as a result of this data. LIME is then prepared to sort out which express aggravations were most important in model evaluations [18].

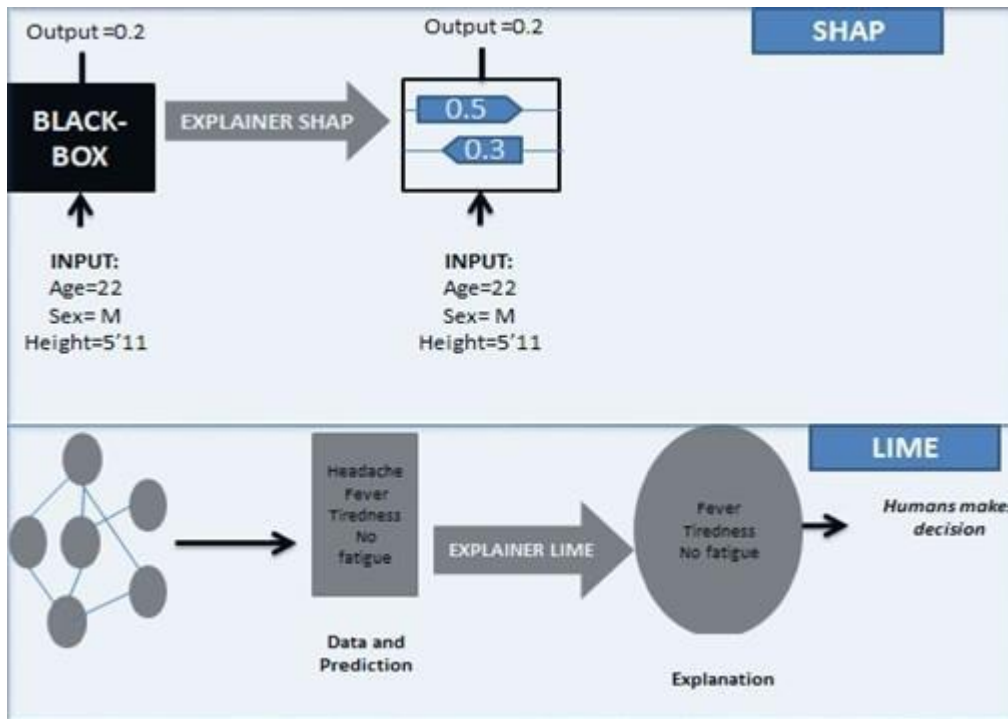


Fig.3. The SHAP and LIME systems are in operation.

CHAPTER 3

Computation OF THE BLACK-BOX

The development of advances in different areas accomplishes a huge load of information that limits human limits, zeroing in on the significance of farsighted and appraisal contraptions that connect better judgment and treatment choices to be accessible meanwhile.

It moreover opens up various assessment open entryways. Computer-based intelligence and significant learning estimations are computations that are altered with PC learning plans for a particular information record and produce novel and ahead of time subtle data

[19].

Calculations for profound learning are a vital piece of the most common way of learning complex capacities. These calculations incorporate portrayal learning strategies that utilize basic and non-direct modules to procure numerous level portrayals [12]. Notwithstanding, current profound learning models have defects that render them insufficient. The nature of the information and the precision of these models are straightforwardly corresponding; gigantic measures of information are expected to prepare the model, which consumes most of the day and essentially affects expectation exactness. While profound learning models are independent, they are exceptionally defenseless to blunders, delivering them problematic and one-sided [20].

The essential explanation that these models are being opposed is an absence of straightforwardness and clarification. Such models are alluded to as 'black box' models since they train themselves utilizing handled information, bringing about a discrete connection between information and result. These models can't make sense of why a result raises worries about the inner usefulness of layers [21]. The customary guide for such models is as the following: information is given as information and handled by computational layers, bringing

about already inconspicuous information at the result. To start, a preparation data file is utilized to foster the model utilizing a particular erudition. Moreover, the learning system speeds up the learning capacity that integrates the information. Following the finish of all aggregate advancement, the machine produces the forecast of the result. The point of convergence mirrors that an expectation is being made without giving any proof to help it. Nonetheless, on account of XAI, another learning interaction fills in as the defense for the result. The extra layer guarantees independent direction is unbiased. It helps with recognizing dataset predispositions, yet in addition, features potential factors that could influence the last expectation.

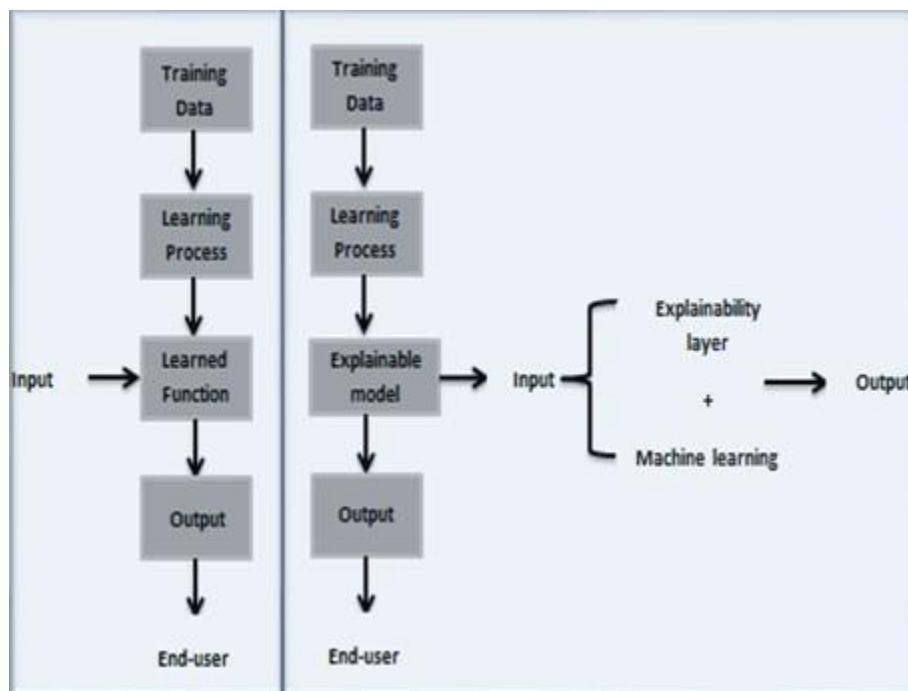


Fig.4. Ordinary AI versus XAI.

The XAI is viewed as a zone of intermingling for an assortment of fields. One of the areas is end-client explanation through sociology, which empowers XAI to foster acumen capacities. Another district in which, where it can exhibit its ability for an explanation, as consistent electronic thinking requires a serious level of a coordinated effort with the client [22].

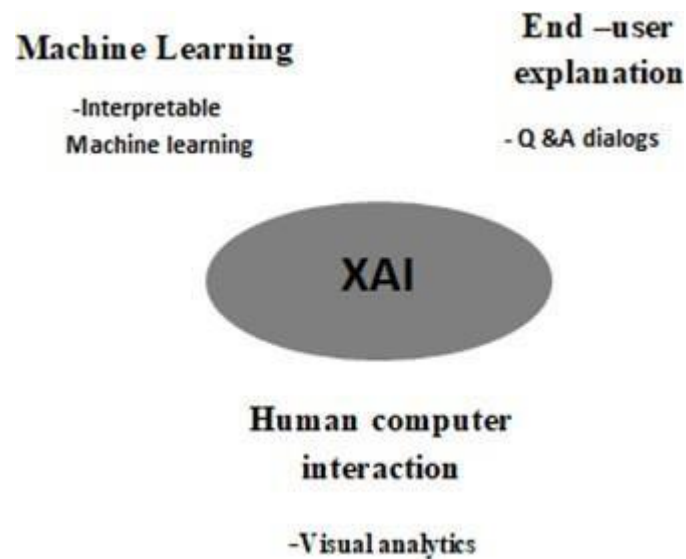


Fig.5. Convergence of XAI with various fields.

CHAPTER 4

Challenges EXPERIENCED BY XAI

Predisposition Issues Data: For the situation of unprejudiced AI, the most customary methodology is to recognize that bad quality information will bring about inferior quality results. The course of the information age, assortment, and handling improves the probability of human predisposition, which brings about slanted and temperamental results. Choice inclination restricts the information and hinders its capacity to be illustrative and various; intrinsic and estimation predisposition compels the accuracy of the information, and the tendency to look for predictable feedback darkens the reviewer's perspective on the information's errors [23].

AI calculations are the consequence of their information; information that habitually goes about as a reflection of society. This unavoidable cultural inclination brings about information that reflects fanaticism and builds up generalizations. The phase of information assortment is not even close to unprejudiced. There are inclinations related to the properties picked, the objective variable picked, and the models picked for incorporation.

Fluctuation in example size is viewed as a huge issue; this causes what is going on in which lacking information brings about unfortunate models with off base results. Also, the information gatherers have blackball control over the informational index's design. Which trait rules while considering the objective still up in the air by the information gatherers' private and expert inclinations. This approach is insignificant to other people who work inside unmistakable structures.

The last stage is to make an AI model utilizing the handled information. This stage is incredibly defenseless to human inclination; steps, for example, including designing can significantly affect

the last AI calculation. The administration of inadequate or missing information can decisively affect the calculation's result [24].

The crude and handled information can assist XAI with working on abilities and uncovering any inconsistencies that might exist. Considering this, XAI can aid the accompanying four regions:

- XAI is fit for recognizing the awkward nature brought about by oversampling.
- XAI can order ascribes that influence both neighborhood and worldwide judgment;
- XAI can feature handling issues that might influence official conclusion making; and
- XAI mechanical assemblies can produce into account the results of client-stamped delicate qualities on model show.

Issues with choices and AI configuration: Biases are one of the elements that impact the calculation's decision; others incorporate the model determination process and the restricting avocation that incidentally presenting parts brings about influencing the result. The AI calculation picked affects the nature of the result. For instance, in relapse models, the connection between factors isn't noticeable. Individual factors are generously esteemed comparable to other accessible factors. In different cases, where inclinations are chosen inside the information, models, for example, Bayesian classifiers are enthusiastically suggested for dealing with this sort of information. Technique choice can be loose. Significantly more extensively, the administered technique is applied to named information; in any case, it is more suitable to involve a solo strategy related to or rather than the directed technique [25]. Creating AI calculations involves capacities for streamlining the calculation. These capacities may inadvertently raise worries about the model's reasonableness. Furthermore, the determination of assessment and usefulness streamlining as far as the model's decency is circumstance subordinate. It is basic to think about sensibility while considering the client's capacity to pick their evaluation metric.

Reasonableness concerns inspire the accompanying assumptions for XAI:

- The XAI can distinguish the effect of AI and advancement choice on general execution.
- While assessing the worldwide show of the resulting computation, the XAI mechanical assemblies think about some norm of fairness.

Issues with the introduction of XAI instruments: The capability of XAI is corresponding to how much it very well may be made sense of. While using the XAI instrument, the client's experience ought to be unessential. The expert uses XAI devices to determine their model's worldwide and neighborhood rehearses. Moreover, the XAI apparatuses ought to be grown explicitly for the crowd for which they are expected. Man-made intelligence-trained professionals, accomplices, and customers all make suspicions about understandability and unprejudiced nature because of the development [24].

CHAPTER 5

APPLICATIONS

Monetary Forecasting: The primary stage in AI producing is include extraction, which is basic for recognizing the model and helping it with classification occupations. This stage is basic and more modern in the monetary space since it extricates highlights from monetary information. With the approach of XAI, a few element choice calculations might be developed in a practical manner in money-related settings [26].

Medical care: The reception of AI in the well-being business has generally brought about the superior direction and the arrangement of the scope of therapy choices. The blend of AI and wearable technology (e.g., Fitbit) has effectively anticipated a client's medical issue by assessing the gadget's wellbeing information. This blend has very strong applications in the domain of medical services yet misses the mark with regards to dealing with the AI-produced black box choices and creates trust concerns. The presentation of XAI and its systems clarifies the reasoning for AI direction. XAI examines well-being information utilizing an AI framework and spotlights the way to accomplishing responsibility, clarity, and following outcomes to build the unwavering quality and reliability of these models[27].

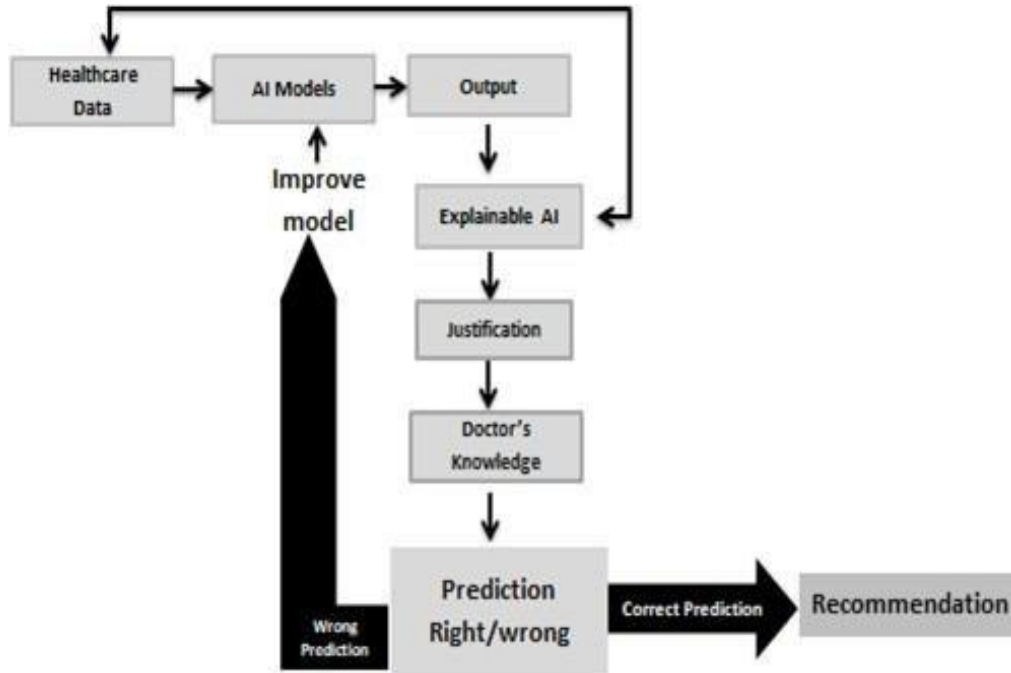


Fig.6.The application of the XAI model in the Healthcare

Industry: The presentation of AI in a few areas has brought about model-initiated inclination. These models raise worries about responsibility and straightforwardness. When applied to industry, the methodology turns out to be financially delicate, underscoring the need for logic. They benefit from the XAI's accentuation of trust and its capacity to decipher replicas at scale [28].

Information Quality: The significance of reasonableness in information quality is basic for conveying inward usefulness while utilizing information-based AI. XAI doesn't simply feature the model; it additionally assesses the information used to assemble the model. Reasonableness guarantees that there is generally an edge for characterizing, perceiving, and making sense of

information defects preceding the structure of the model, as well as giving the remedial activity. As a rule, reasonableness can be utilized to legitimize depicting information as far as AI and assessing or measuring it [29].

CHAPTER 6

LITERATURE REVIEW:

The Toxic Substances Control Act (TSCA) manages around 84,000 mixtures, essentially all of which have not been read up for formative neurotoxicity. The developing human cerebrum is especially helpless against harmful openings, and the assessed expenses of formative neurotoxicity are significant. In this manner, fast, minimal expense, and precise methodologies for evaluating formative neurotoxicity are required direly [30].

Current methodologies remember for Vivo creature examinations and evaluations of essential cell and tissue societies from creatures and people. These strategies are restricted by time limitations, costs, the shortage of essential cells and tissues, and their irrelevance to human physiology. Because of these snags, drug endorsement rates have diminished in spite of expansions in innovative work burning through. Human pluripotent foundational microorganisms can help with resolving these issues by offering an adaptable wellspring of clinically valuable human cells at a sensible expense [30].

The National Institutes of Health (NIH) sent off the Microphysiological Systems Program in 2012 as a team with the Defense Advanced Research Projects Agency (DARPA) and the United States Food and Drug Administration (FDA), fully intent on creating human tissue chips containing bioengineered tissue models that copy human physiology. The reason for this program is to utilize these chips to make forecasts about the wellbeing and adequacy of expected

meds. Schwartz et al. utilized quality articulation information from these manifestations to create 3D models of developing human brain tissue from an assortment of cell types and prepared AI models to foresee formative neurotoxicity. The 3D model is prepared to do exactly identify risky substances and showing the capacity to summarize significant natural frameworks, demonstrating the opportunities for extra information. In any case, this 3D model is intrinsically intricate and requires a lot of mastery and actual work to effectively create and convey [30].

METHODOLOGY:

- **DATA COLLECTION**

The data has been acquired from GitHub, DevTox2D provides Code and handled information from a correlation concentrated on contrasting 3D human tissue models and 2D human tissue models for foreseeing formative neurotoxicity. Data contains 180 rows and 11 columns. These are the datasets of quality articulation profiles got from 3D and 2D human neural tissue models presented to an assortment of substances. The datasets are communicated in records per million (TPM), with each line signifying a solitary example and every section meaning a quality. Each example addresses tissue openness to a solitary harmful or non-poisonous compound at a single exposure length. Test names are in the primary segment and recognize the compound and openness length, with contrasting naming configurations for the 2D and 3D datasets.

Expression_2D:

The example name D39-C15 in the 2D tissue model dataset compares to the non-toxic substance c15 following 39 days of openness. D27Hi-4 is the example name for the toxic synthetic t4 after 27 days of openness.

Expression_3d:

The example names in the 3D tissue model assortment follow a reliable example. Notwithstanding the two organic copies of each example in the 2D dataset, the 3D dataset contains two natural imitates of each example. For example, the sample name d2c15b compares to the non-toxic synthetic c15 at a two-day openness time point, organic imitates two. The example d7t4a is an agent test of the unsafe synthetic t4 following seven days of openness, a natural repeat one. example d7t4a is an agent test of the unsafe synthetic t4 following seven days of openness, natural repeat one.

FEEDING DATASET ON MACHINE LEARNING:

The datasets were haphazardly partitioned into preparing and testing sets in a 80:20 proportion. Because of the developing fame of AI strategies like SVMs, KNNs, and DL in disciplines, for example, omics information examination, arrangement information investigation, biomedical imaging, and sign handling we decided to lead AI on our datasets. Three XGBoost models for arrangement were prepared to utilize the preparation sets. The XGBoost strategy (Extreme Gradient Boosting) is an AI method that utilizes choice trees to enhance execution through an interaction called supporting. It has constantly beaten the heft of other AI strategies, including calculated relapse, the irregular woods model, and traditional choice trees, since its presentation. XGBoost systems are accessible for an assortment of scripting languages, most outstandingly Python, and they incorporate pleasantly with the famous scikit-learn AI structure that Python

information researchers use. Following the organization of the XGBoost AI classifier employing the Scikit-learn tool compartment to the datasets, the models' presentation was assessed utilizing the testing sets. Models were assessed concerning their disarray network and the model's precision when applied to the test set [31] .

APPLYING XAI ON TRAINED ML:

The Python SHAP (Shapley Additive exPlanations) library was utilized to perform XAI investigation on the XGBoost models that had been prepared. Exploring model choices with the assistance of an XAI investigation can assist with finding factors that lift model trust in its expectations. For every quality, the normal/mean outright worth across all examples was utilized to work out the worldwide importance. This data was then given to the bar plot work, which built a worldwide component significant plot. We had the option to recognize the main qualities by utilizing this element significance plot.

CONFUSION MATRIX:

While calculating sensitivity, the specificity TP, FP, TN and FN values assists to reach the conclusion. It is NxN matrix that determines the performance of the model for classification. The components of the confusion are used to observe three significant boundaries named precision, sensitivity, and specificity. The expectation of classes for the information in a grouping issue depends on tracking down the ideal limit between classes. In light of the upsides of precision, sensitivity, and specificity one can track down the ideal limit.

CHAPTER 7:

RESULT

Finally, our model has achieved an accuracy of 86%, assisting in determining the impact of genes on our model's performance. ID4, WWC1, SRXN1, BMP7, GRIN3A, NCF2, DDN, F13A1, RNF207, RNASE1, SRPX, KDSR, CELF4, HIST1H3H, MMP12, SRSF6, PRKCSH, TAP1, DALRD3 are the main genes in all datasets, as per the going with SHAP summary charts, and their consequences for model forecasts are both huge and positive. Utilizing the dataset's exactness, we can test our projected genes' exhibition and perceive what it means for our AI model. Beneath, you can see the confusion matrix.

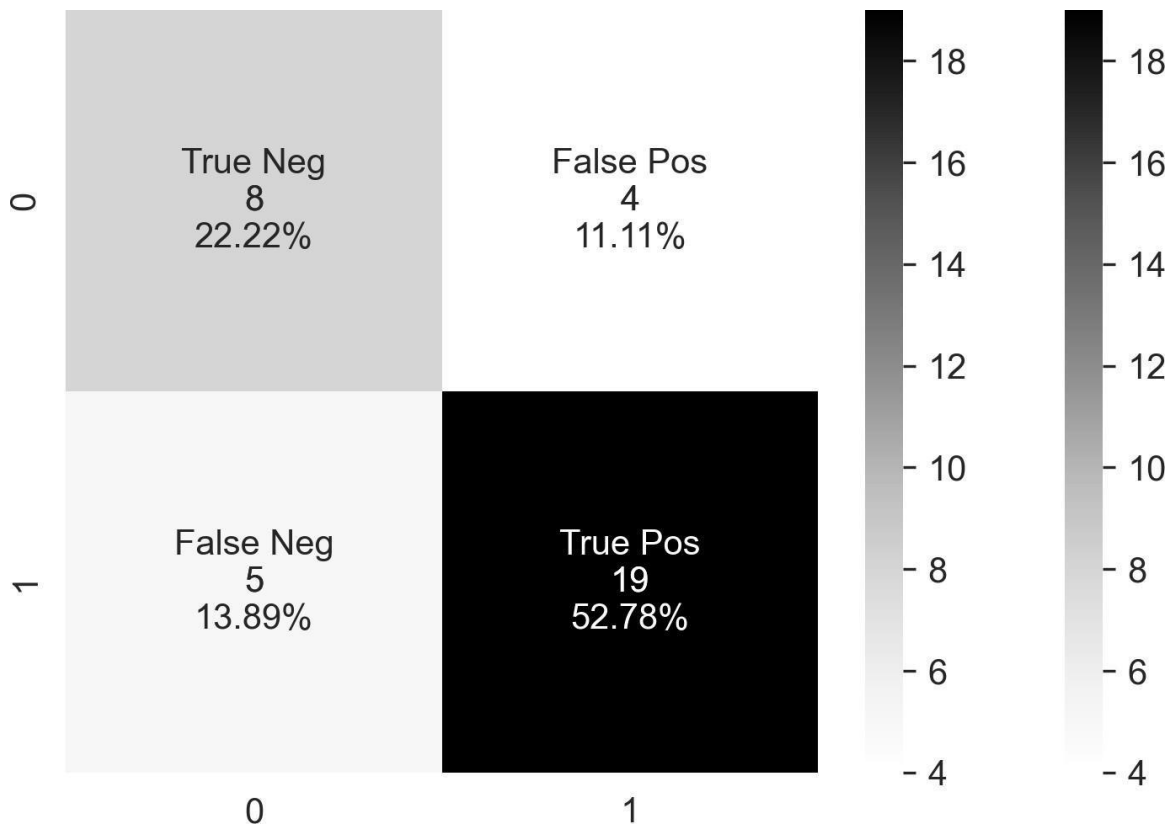


Fig 7. The quantity of True positive cases is shown by the grey squares in the grid (TP), False-positive (FP) and False negative (FN) models are addressed by dark squares in the grid, while true negative (TN) events are addressed by white squares in a similar way.

Furthermore, the SHAP synopsis plots were made and used to show the positive and negative associations between the indicators and the objective quality . It's obvious from these SHAP outline graphs that: It is feasible to see whether a quality smaller affects the model result by taking a gander at its level situation; the shade of the line shows if a quality's impact is extensive (in red) or negligible (in blue) for an offered perspective. It altogether affects the quality positioning if the 'ID4' quality is available in high overflow. Utilizing the X-axis, the "positive" impact is portrayed by the red tone.

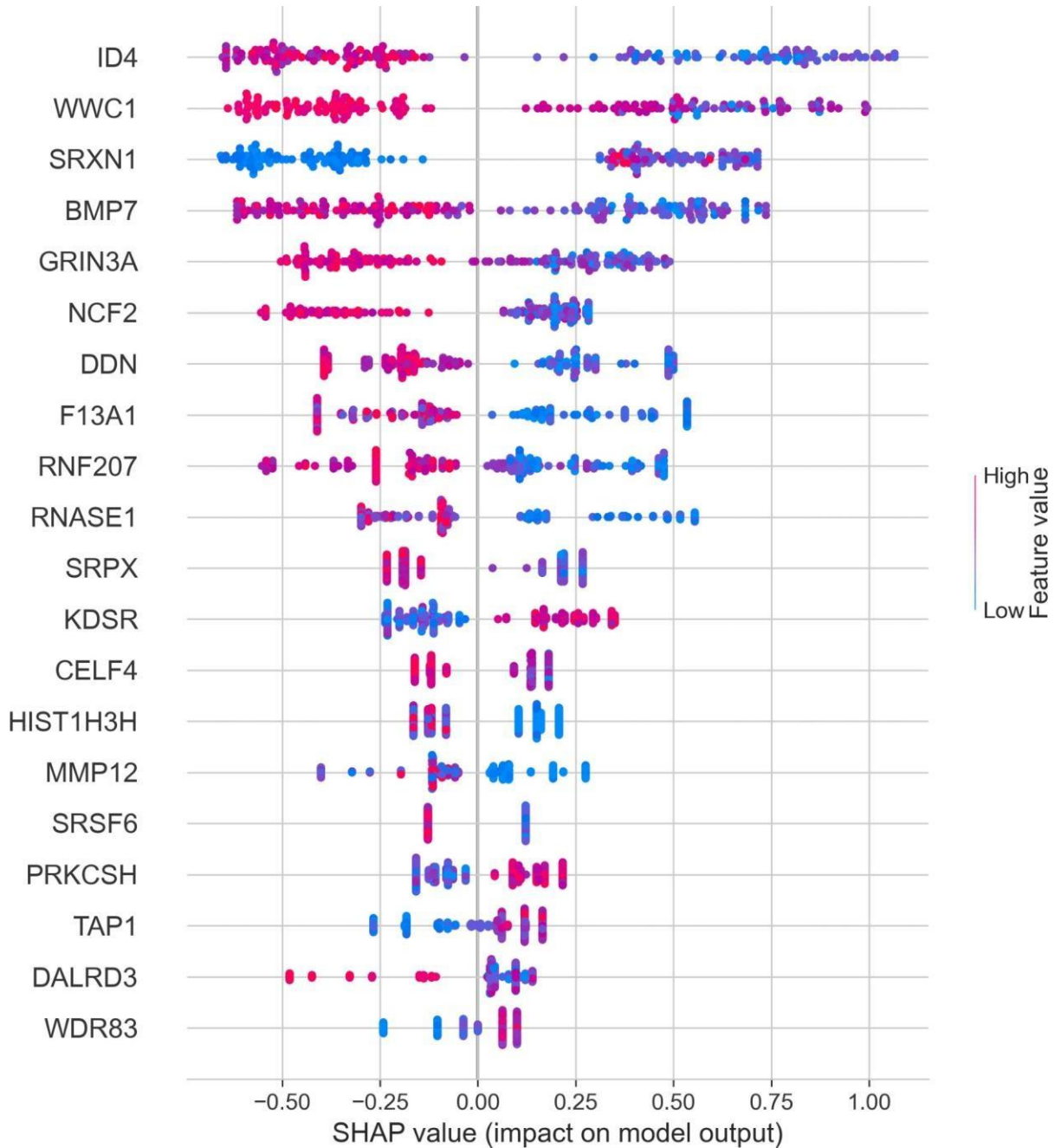


Fig 8. The plot of the SHAP synopsis shows the main qualities and their effect. The y-axis is utilized to rank the significance of every quality element on the y-hub. Shown on the x-axis are the consequences of a quality's impact on the model's result: expanded or decreased forecast. Red demonstrates a quality's impact is genuinely critical, while blue shows no impact by any means.

The prescient force of the ML model is helped by qualities towards the highest point of the tree, as opposed to those at the base. Fig. shows bar plots that show the main genes at the top and the most uncritical genes at the base.

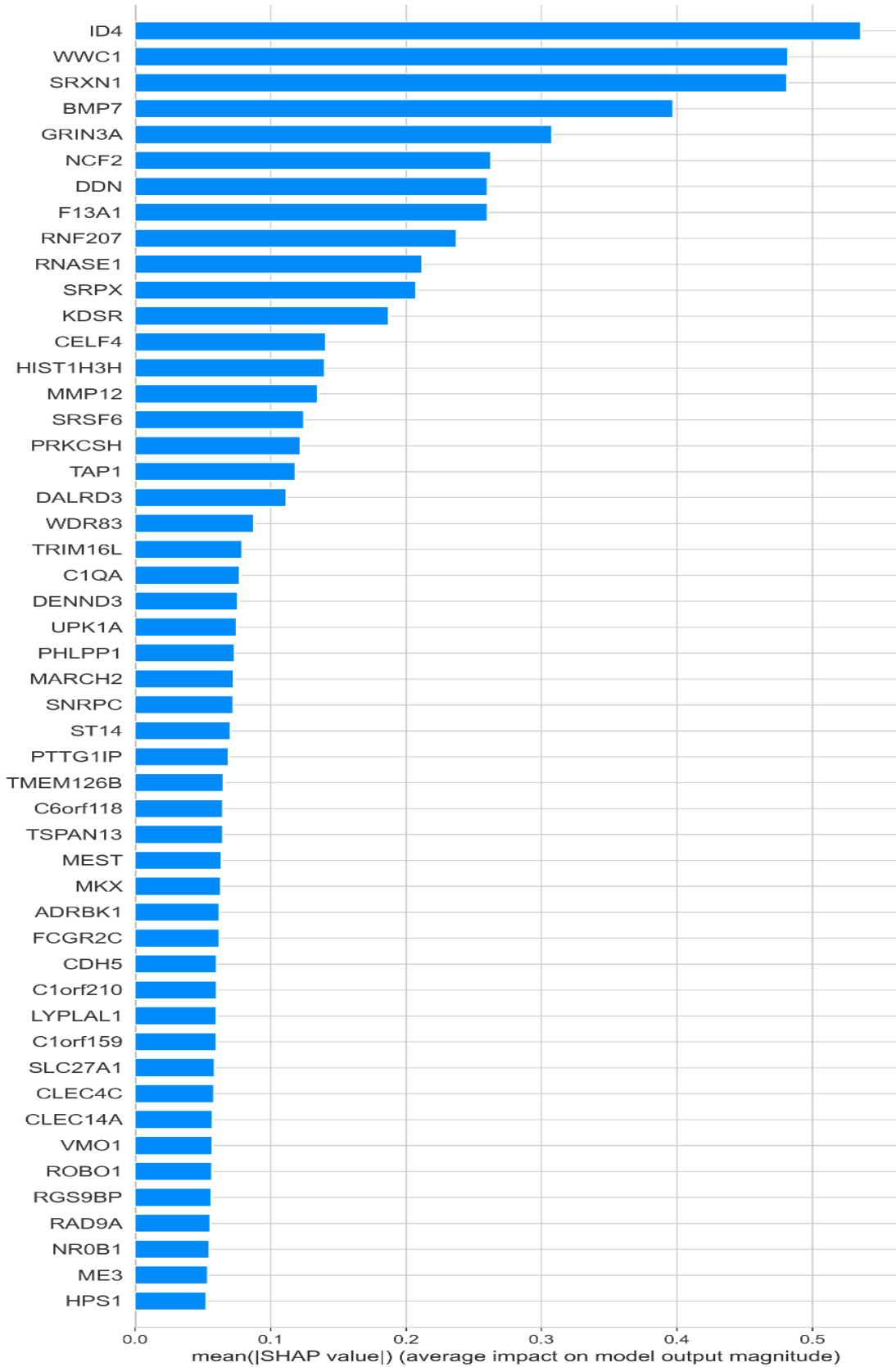


Fig 9. From the given bar plot, genes of most noteworthy significance were driven by carrying out SHAP values on the prepared models, showing ID4, WWC1, SRXN1, BMP7, GRIN3A, NCF2, DDN, F13A1, RNF207, RNASE1, SRPX, KDSR, CELF4, HIST1H3H, MMP12, SRSF6, PRKCSH, TAP1, DALRD3.

VALIDATION:

Our model has identified the top genes which have impacted the model’s accuracy. When these genes are extracted and again fed as input, the accuracy has increased from 86% to 91.67%. Hence, this proves that our model has successfully identified the role of key genes that affected the model’s accuracy and performance.

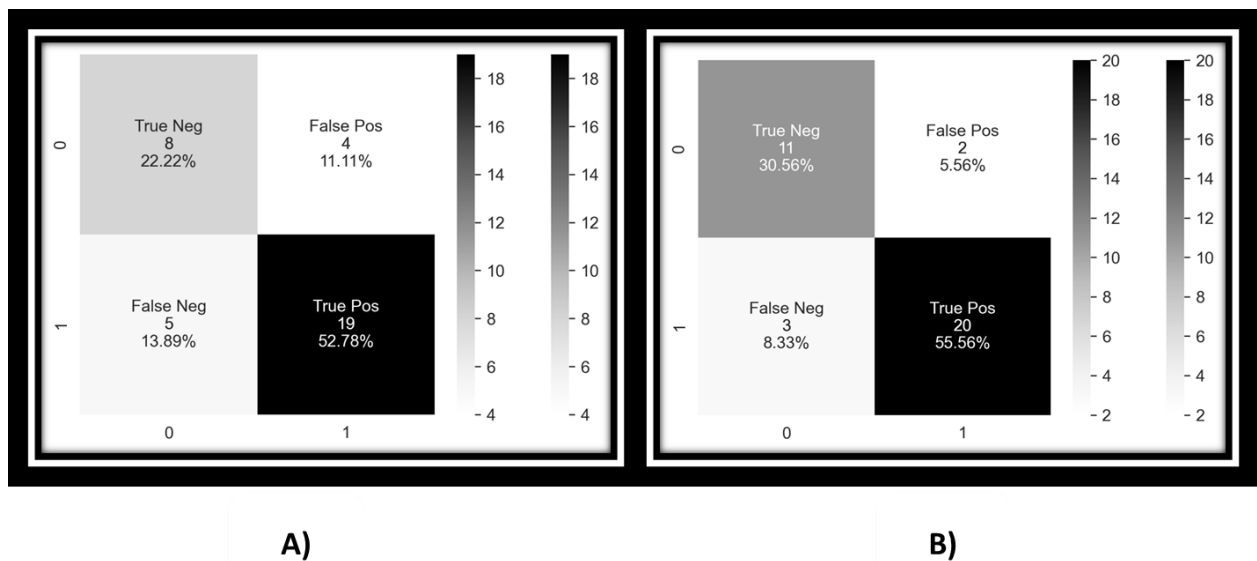


Fig 10. Confusion matrix (A) shows the accuracy of the model containing all genes, whereas the confusion matrix (B) shows improved accuracy due to previously identified key genes.

A)

```
In [17]: import xgboost as xgb
```

```
In [30]: xgb_mod=xgb.XGBClassifier(random_state=42) # build classifier  
xgb_mod=xgb_mod.fit(X_train,y_train)
```

```
In [31]: from sklearn.metrics import confusion_matrix, accuracy_score  
y_pred = xgb_mod.predict(X_test)  
y_pred_prob = xgb_mod.predict_proba(X_test)  
AK_probs = y_pred_prob[:,1]  
  
# Performance  
accuracy = accuracy_score(y_test, y_pred)  
  
print("Accuracy: %.2f%%" % (accuracy * 100.0))
```

Accuracy: 86.11%

B)

```
In [37]: import xgboost as xgb
```

```
In [53]: xgb_mod=xgb.XGBClassifier(random_state=42) # build classifier  
xgb_mod=xgb_mod.fit(X_train,y_train)
```

```
In [54]: from sklearn.metrics import confusion_matrix, accuracy_score  
y_pred = xgb_mod.predict(X_test)  
y_pred_prob = xgb_mod.predict_proba(X_test)  
AK_probs = y_pred_prob[:,1]  
  
# Performance  
accuracy = accuracy_score(y_test, y_pred)  
  
print("Accuracy: %.2f%%" % (accuracy * 100.0))
```

Accuracy: 91.67%

Fig 11. (A) shows the accuracy of 86% in which the model selected the key genes, whereas (B) shows the high accuracy of 91.67% when the key genes are again fed to the model as input.

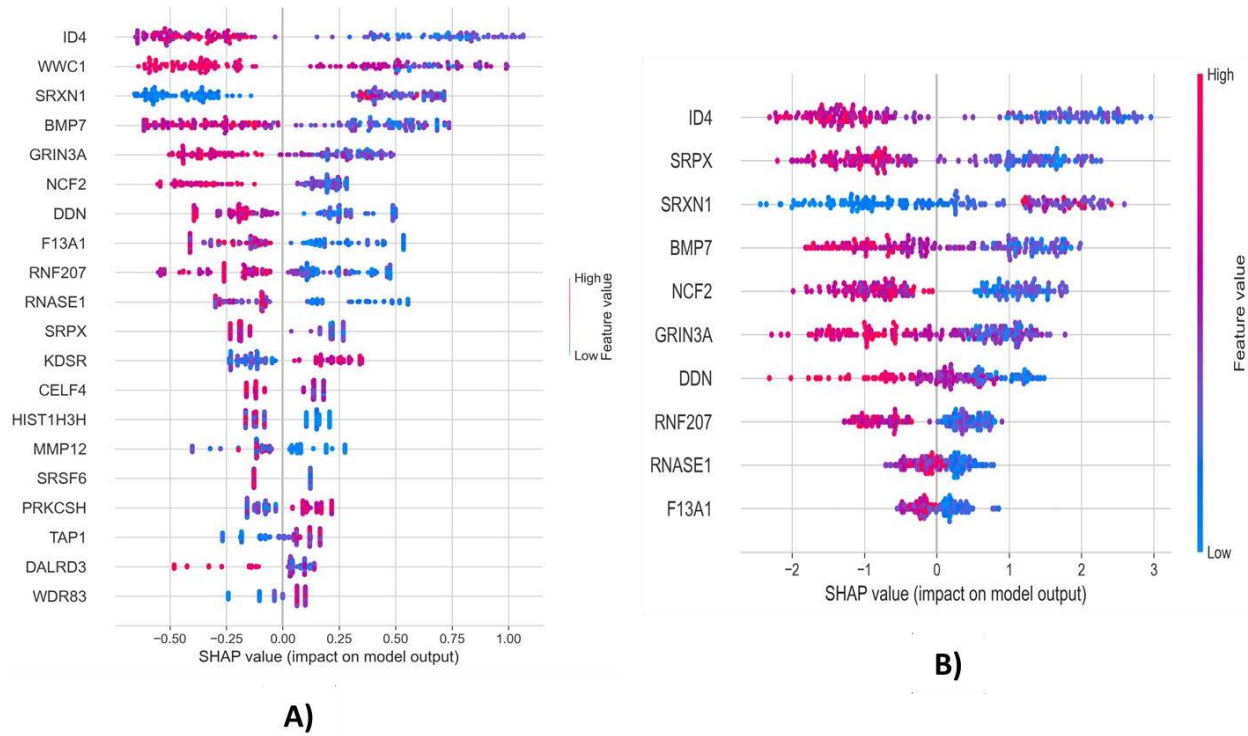


Fig 12. SHAP Values of pervious model and validation model

The biological significance of the significant genes that have been found:

ID4:

This gene gives guidelines for making a protein that has a place with the inhibitor of the DNA binding (ID) protein family. Subsequently, the produced protein misses the mark on the ability to bind DNA and instead controls gene expression by binding to and inhibiting transcription factors with the basic helix-loop-helix structure. This protein has been embroiled in the control of a wide scope of cell exercises, incorporating those associated with the making of growths and the movement of disease. ID4 (Inhibitor Of DNA Binding 4, HLH protein) is a protein-coding quality that is found in the human genome. Id proteins, a group of helix-circle helix (HLH) proteins that remembers four individuals from Id1 to Id4 for mammalian cells, are fundamental for the guideline of cell development, separation, senescence, cell cycle movement, and

expanding angiogenesis and vasculogenesis, as well as the capacity of cells to relocate quicker than typical. Whenever Alzheimer's illness (AD) uncovers itself in the grown-up populace, it shows proof of mental deterioration, conduct anomalies, and utilitarian hindrance. Alzheimer's infection is the most pervasive neurodegenerative disease in the grown-up populace. The aggregation of amyloid-beta peptides (As), which are the essential parts of the feeble plaques shown in Alzheimer's sickness minds, contributes incredibly to the course of the disease. A developing assortment of information recommends that strange cell cycle reemergence may assume a basic part in the mortality of neurons created by A. Various flagging go-betweens, including Id1, hypoxia-inducible component 1 (HIF-1), cyclin-dependent kinases-5 (CDK5), and sonic hedgehog (Shh), have as of late been found to be associated with A-initiated cell cycle reemergence in postmitotic neurons. Besides, Id1 and CDK5/p25 have been displayed to adversely manage the articulation and action of one another. Accordingly, Id proteins might have expected restorative purposes in Alzheimer's sickness [32].

SRXN1:

It is accepted that oxidative stress, which brings about protein oxidation, is a contributing element to mental ailments, like Parkinson's infection, Alzheimer's sickness, and stroke. As well as being an endogenous cancer prevention agent protein, sulfiredoxin-1 (Srxn1) additionally has neuroprotective properties. Srxn1 can shield PC12 cells from oxidative pressure created by H₂O₂ and is associated with the capacity of Prdxs. Srxn1 has been displayed to have a defensive capacity against oxidative harm and can possibly be utilized as an objective for neuroprotective mediation in oxidative stress studies [33].

BMP7:

This quality encodes an emitted ligand of the TGF-beta (transforming growth factor-beta) class of proteins, which is associated with cell advancement and separation. Ligands from this family are tied to an assortment of TGF-beta receptors, causing the enrollment and initiation of record factors from the SMAD family, which control quality records. It is the proteolytic handling of the encoded preproprotein that outcomes in the arrangement of every part of the disulfide-connected homodimer, which is engaged with the advancement of the bone, kidney, and brown fat tissue. Moreover, this protein has been displayed to invigorate ectopic bone development in creature models and to support the mending of breaks in people [34].

NCF2:

To keep up with great well-being, it is important to keep an appropriate redox balance between the development and evacuation of receptive oxygen species (ROS). An interruption in the climate, for example, relentlessly expanded ROS levels, will bring about oxidative pressure and the advancement of the disease, while a deficient measure of ROS age will be negative to wellbeing. It is for the most part due to inactivating varieties in the qualities encoding for NADPH oxidase buildings that ROS creation is diminished or even totally disposed of. Specifically, hereditary varieties (CYBB, CYBA, NCF1, NCF2, NCF4) that actuate an imperfection in phagocyte Nox2 oxidase action have been distinguished as an immediate reason for persistent granulomatous illness (CGD), an inherited immunological condition .

DDN:

Represses DNA-restricting record factor movement and RNA polymerase II-explicit DNA restricting action, as well as DNA restricting action intended for the RNA polymerase II cis-regulatory district succession. It is anticipated that RNA polymerase II will work either

upstream of or inside the positive control of the record. Cell projection and the cytoplasm are the areas of this gene[35].

F13A1:

Parts of factor XIII subunit A (FXIII-A function)'s other than hemostasis are likewise significant. It is additionally tracked down intracellularly in different human cells and can be utilized as an indicative marker in the treatment and analysis of a wide scope of dermatological ailments, going from fiery states to harmful growths. Utilizing this audit, we need to give a direction on the still challenging translation of dermal cell types that express FXIII-A, as well as an evaluation of the recently reported processes that lead to their collection under physiological and obsessive circumstances in the skin of the human [36].

RNF207:

Takes into consideration the limiting of Hsp70 proteins, the limiting of chaperones, and the limiting of transmembrane carriers to the plasma film. Takes an interest in the positive guideline of deferred rectifier potassium channel movement, the positive guideline of quality articulation, and the positive guideline of voltage-gated potassium divert action in ventricular heart muscle cells that are engaged with the repolarization of the activity capability of the cardiovascular muscle cell. The perinuclear region of the cytoplasm is where this protein is found [37].

CHAPTER 8

CONCLUSION

Current methodologies for toxicity screening are simply excessively sluggish and costly to be utilized to look at all novel or inadequately figured out compound openings in a thorough way. The utilization of these models is nearly expected to go on for a long time to come, albeit the utilization of high-throughput screening techniques has a huge guarantee. The consequences of AI-established yield are by and large relative to the requirement for XAI (abundance AI). While utilizing XAI, the viewpoints that can change the final result might be accentuated since it offers the administrator gainful information on models through which the future utilization of such a model can be chosen further. XAI is based on top of various structures, the most generally utilized of which being SHAP and LIME, in spite of the fact that there are others. Nonetheless, the advancement of each innovation is relative to the number of counter-challenges it faces. XAI has gone up against various impediments, incorporating issues with one-sided information, which brings about segregation in the result, and trouble with creating calculations for an ML model, which brings about an inevitable inventive inclination in the model. This work researched the utilization of XAI to track down competitor genes, explicitly ID4, WWC1, SRXN1, BMP7, GRIN3A, NCF2, DDN, F13A1, RNF207, RNASE1, SRPX, KDSR, CELF4, HIST1H3H, MMP12, SRSF6, PRKCSH, TAP1, DALRD3. All these genes have a great impact on the model's performance in predicting neurotoxicity.

REFERENCES:

- [1] K. Alikhademi, B. Richardson, E. Drobina, and J. E. Gilbert, “Can Explainable AI Explain Unfairness? A Framework for Evaluating Explainable AI,” Jun. 2021, Accessed: Dec. 28, 2021. [Online]. Available: <http://arxiv.org/abs/2106.07483>
- [2] R. K. E. Bellamy *et al.*, “AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias,” Oct. 2018, Accessed: Dec. 28, 2021. [Online]. Available: <http://arxiv.org/abs/1810.01943>
- [3] J. Dodge, Q. Vera Liao, Y. Zhang, R. K. E. Bellamy, and C. Dugan, “Explaining models: An empirical study of how explanations impact fairness judgment,” *International Conference on Intelligent User Interfaces, Proceedings IUI*, vol. Part F147615, pp. 275–285, 2019.
- [4] Wojciech Samek, Thomas Wiegand, and Klaus-Robert Müller, “Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models.,” *arXiv preprint arXiv:1708.08296 (2017)*, 2017.
- [5] N. Haefner, J. Wincent, V. Parida, and O. Gassmann, “Artificial intelligence and innovation management: A review, framework, and research agenda☆,” *Technological Forecasting and Social Change*, vol. 162, p. 120392, Jan. 2021, doi: 10.1016/J.TECHFORE.2020.120392.
- [6] Shane T. Mueller and Elizabeth S. Veinott, “Principles of Explanation in Human-AI Systems,” 2021.
- [7] Yann LeCun and Yoshua Bengio, “Deep learning,” May 2015.
- [8] Mervin Jeff, “When will we get there?,” *Science V*, 2017.

- [9] Jenkins C and Lopresti D, “Next Wave Artificial Intelligence: Robust, Explainable, Adaptable, Ethical, and Accountable,” 2020.
- [10] Kristian Kersting, “Rethinking Computer Science Through AI,” Dec. 2020.
- [11] KAI-FU LEE, “The Four Waves of A.I.,” Nov. 23, 2018.
- [12] Y. Lecun, Y. Bengio, and G. Hinton, “Deep learning,” *Nature*, 2015.
- [13] Feiyu Xu and Hans Uszkoreit, “Explainable AI: A Brief Survey on History, Research Areas, Approaches and Challenges,” Sep. 2019.
- [14] Holger Hoos and Kristian Kersting, “The third wave of artificial intelligence,” Dec. 2020.
- [15] “Artificial Intelligence Tech Will Arrive in Three Waves.” <https://futurism.com/artificial-intelligence-tech-will-arrive-in-three-waves> (accessed Dec. 28, 2021).
- [16] Danding Wang and Qian Yang, “Designing Theory-Driven User-Centric Explainable AI,” *CHI '19: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, May 2019.
- [17] M. Chromik, “reSHAPe: A Framework for Interactive Explanations in XAI Based on SHAP,” *European Society for Socially Embedded Technologies (EUSSET)*, 2020.
- [18] J. Dieber, “Why model why? Assessing the strengths and limitations of LIME,” Nov. 2020.
- [19] Došilović FK, “ Explainable artificial intelligence: A survey. ,” *41st International convention on information and communication technology, electronics and microelectronics (MIPRO). IEEE*, 2018.
- [20] Schnack H, “interpretability in machine learning: From measurements to features. In: *Machine Learning. , Academic Press*, 2020.

- [21] Samek W and Wiegand T, “Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models,” 2017.
- [22] E. Dağlarlı, “Explainable Artificial Intelligence (xAI) Approaches and Deep Meta-Learning Models,” *Advances and Applications in Deep Learning*, Jun. 2020, doi: 10.5772/INTECHOPEN.92172.
- [23] J. Dodge, Q. Vera Liao, Y. Zhang, R. K. E. Bellamy, and C. Dugan, “Explaining models: An empirical study of how explanations impact fairness judgment,” *International Conference on Intelligent User Interfaces, Proceedings IUI*, vol. Part F147615, pp. 275–285, 2019.
- [24] K. Alikhademi, B. Richardson, E. Drobina, and J. E. Gilbert, “Can Explainable AI Explain Unfairness? A Framework for Evaluating Explainable AI,” Jun. 2021, Accessed: Dec. 29, 2021. [Online]. Available: <http://arxiv.org/abs/2106.07483>
- [25] E. Toreini, M. Aitken, K. Coopamootoo, K. Elliott, C. G. Zelaya, and A. van Moorsel, “The relationship between trust in AI and trustworthy machine learning technologies,” *FAT* 2020 - Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 272–283, Jan. 2020,
- [26] Salvatore Carta, “Explainable AI for Financial Forecasting”.
- [27] U. Pawar, D. O’Shea, S. Rea, and R. O’Reilly, “Explainable AI in Healthcare,” *2020 International Conference on Cyber Situational Awareness, Data Analytics and Assessment, Cyber SA 2020*, Jun. 2020.
- [28] KrishnaGade and SahinCemGeyik, “Explainable AI in Industry,” *the 25th ACM SIGKDD International Conference*, 2019.
- [29] L. Bertossi and F. Geerts, “Data Quality and Explainable AI,” *Journal of Data and Information Quality*, vol. 12, no. 2, May 2020.

- [30] F. Kuusisto, V. Santos Costa, Z. Hou, J. Thomson, D. Page, and R. Stewart, “Machine learning to predict developmental neurotoxicity with high-throughput data from 2D bio-engineered tissues,” *Proc Int Conf Mach Learn Appl*, vol. 2019, p. 293, Dec. 2019, doi: 10.1109/ICMLA.2019.00055.
- [31] J. Meena and Y. Hasija, “Application of explainable artificial intelligence in the identification of Squamous Cell Carcinoma biomarkers,” *Computers in Biology and Medicine*, vol. 146, p. 105505, Jul. 2022, doi: 10.1016/J.COMPBIOMED.2022.105505.
- [32] S. der Chen, J. L. Yang, Y. C. Lin, A. C. Chao, and D. I. Yang, “Emerging Roles of Inhibitor of Differentiation-1 in Alzheimer’s Disease: Cell Cycle Reentry and Beyond,” *Cells* 2020, Vol. 9, Page 1746, vol. 9, no. 7, p. 1746, Jul. 2020, doi: 10.3390/CELLS9071746.
- [33] Q. Li, S. Yu, J. Wu, Y. Zou, and Y. Zhao, “Sulfiredoxin-1 protects PC12 cells against oxidative stress induced by hydrogen peroxide,” *J Neurosci Res*, vol. 91, no. 6, pp. 861–870, Jun. 2013, doi: 10.1002/JNR.23218.
- [34] “BMP7 bone morphogenetic protein 7 [Homo sapiens (human)] - Gene - NCBI.” <https://www.ncbi.nlm.nih.gov/gene/655> (accessed May 03, 2022).
- [35] F. Dunér *et al.*, “Dendrin expression in glomerulogenesis and in human minimal change nephrotic syndrome,” *Nephrology Dialysis Transplantation*, vol. 23, no. 8, pp. 2504–2511, Aug. 2008, doi: 10.1093/ndt/gfn100.
- [36] L. Paragh and D. Törocsik, “Factor XIII Subunit A in the Skin: Applications in Diagnosis and Treatment,” *BioMed Research International*, vol. 2017, 2017, doi: 10.1155/2017/3571861.
- [37] K. Roder *et al.*, “RING finger protein RNF207, a novel regulator of cardiac excitation,” *Journal of Biological Chemistry*, vol. 289, no. 49, pp. 33730–33740, Dec. 2014, doi: 10.1074/jbc.M114.592295.

PAPER NAME

Thesis Work

AUTHOR

Nakul Tanwar

WORD COUNT

4728 Words

CHARACTER COUNT

26962 Characters

PAGE COUNT

27 Pages

FILE SIZE

1.0MB

SUBMISSION DATE

Apr 30, 2022 1:41 PM GMT+5:30

REPORT DATE

Apr 30, 2022 1:42 PM GMT+5:30**● 3% Overall Similarity**

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

- 1% Internet database
- 2% Publications database
- Crossref database
- Crossref Posted Content database
- 0% Submitted Works database

● Excluded from Similarity Report

- Bibliographic material

● 3% Overall Similarity

Top sources found in the following databases:

- 1% Internet database
- Crossref database
- 0% Submitted Works database
- 2% 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	Nakul Tanwar, Yasha Hasija. "Explainable AI; Are we there yet?", 2022 I...	2%
	Crossref	
2	assets.researchsquare.com	<1%
	Internet	
3	UW, Stevens Point on 2022-04-28	<1%
	Submitted works	
4	dashboard-pmb.nurulfikri.ac.id	<1%
	Internet	

Explainable AI; Are we there yet?

Nakul Tanwar
Department of Biotechnology
Delhi Technological University
Delhi, India
tbret41@gmail.com

Yasha Hasija*
Department of Biotechnology
Delhi Technological University
Delhi, India
yashahasija06@gmail.com

*Corresponding Author

Abstract— Artificial intelligence (AI) and its models have made a significant impact when it comes to performing tasks that limit human capabilities; this not only reduces human efforts but also detaches human errors. The introduction of such models in various fields results in output which subsequently ameliorates laborious chores such as data processing and other prediction tasks. However, these models produce compound outputs which can be baffling for humans and makes these models vague in terms of providing the reason behind the output. The opaqueness in the internal functionality in the models makes them ‘black-box’ which not only raises a question in terms of the trust but also increases the chances of biasness in the model. When black-box models are applied in real-life problems where transparency is the biggest aspect, the end-user will always be in a dilemma whether to trust the decision produced by these models or not. In recent years, researchers have put a lot of effort into finding a tool that eclipses all these issues and came up with explainable artificial intelligence (XAI) that helps in justifying the output of such models, but with the advancement, there will be some counter challenges always. The extent of the paper is to feature the urgency of XAI to overshadow the downsides of these models and to inspect the impact of XAI in various sectors. Afterwards framework of XAI and challenges faced by XAI are also being discussed.

Keywords— Artificial Intelligence, Black-box, Explainable artificial intelligence, Challenges, Transparency

I. INTRODUCTION

In today’s time machine learning and deep learning both are the forefront algorithms that not only produce astounding outcome but also remove the human role when it comes to handling, storing, and processing the data [1]. The main agenda of these models is to increase precision, accelerate processes and perform decision-making tasks which increase the human dependency on these models. But these models are deficient in terms of transparency, impartiality, and explanation which make them irresolute in applying them in actual life problems [2]. XAI comes with a resolution to provide the internal functionality of the layers that can make these models trustworthy and lucid. XAI is designed in such a way that it provides a shred of proper evidence that supports the output and highlights features that can change the final decision. As this tool is very new, its potential role has to be explored in different domains which include forecasting, healthcare, and industry.

The paper is categorized as follows: segment 2 describes related work and principles of XAI, segment 3 presents the hierarchy of AI, segment 4 explains the framework of XAI, segment 5 summarizes the black-box algorithm, segment 6

explores the challenges of XAI. Several applications of XAI and conclusion with future perspectives are being summed up in segments 7 and 8 respectively.

II. RELATED WORK

The term ‘explainable’ or ‘explainability’ is referred to the potential of the model to justify its outcome. Explainability also guarantees the model to act under canon and secures the accuracy of models [3]. Here, it can further describe the utter and precise depiction of output generated by a model. Explainability is imaginably divided into local (for a solitary occasion of a decision), or global (for grasping the model's decision-prediction algorithm). The arising importance of explainability highlights the urgency of tools that help humans with understanding the conduct of black-box models. Such tools frequently allude to (XAI) [4].

The introduction of advanced technologies leads to the production of a large amount of data. This data is managed and processed by AI and its technologies (machine learning and deep learning) that make humans more dependent on such models [5]. This dependency makes humans in a rightful position to know the exact intention behind any outcome. The focus of XAI is to guarantee that the end-user is receiving the reason behind the outcome which boosts these models and makes them more reliable.

The XAI is enfolded in the subsequent principles:

A. Explanation

The foremost priority of XAI is to provide complete corroboration to reinforce the outcome. There are five sorts of clarifications in AI—for clients, to acquire trust in the public eye, to meet administrative and consistency prerequisites, to create AI models with AI calculations, and for framework proprietors [6].

B. Meaningful

This principle implies that XAI should be purposeful to the operator and administration to grasp an explanation in various ways as well as deal with serious consequences regarding various degrees of inquiry [6].

C. Precision

Precision is a very important aspect for XAI; this does not only explain the reason behind the outcome but also provides an insight into the decision-making of a real problem in a jiffy. This will also build trust in the eyes of end-user while using the AI models [6].

D. Knowledge Curtailment

Every technology has its limits, applying this principle XAI should use knowledge limits regarding training data that will help in preventing the production of unsuitable insights [6].

The main focus of this paper is to inscribe the articles from several years to 2021 to highlight the current situation of XAI. It also comes up with different stages and components which are elite that provide robustness to the conclusion drawn from several articles and researches. We believe this learning will give a valuable angle to analysts; it outfits a total outline of this complex and always advancing topic.

III. THE HIERARCHY OF AI

Advancing algorithms push AI that can act and react by carrying out laborious tasks such as classification, recognition, recommendation, interpretation, and configuration [7]. The introduction of AI in the healthcare/biomedical sector not only improves judgment but also provides various options for treatment. But the full potential of such technologies still has to put place that raises the question regarding the current status of AI [8]. In consideration to this, the defense advanced research project agency (DARPA) forked out the ‘three waves of AI’ that provides the answer for estimation of AI and summaries the proficiency and drawbacks [9].

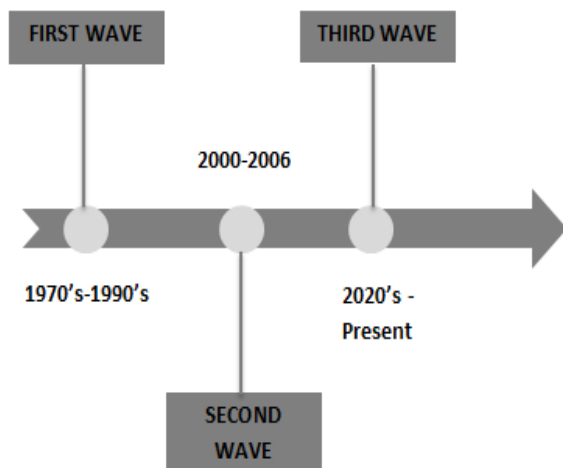


Fig.1. Timeline representation of three waves of AI

A. The First Wave: The Indigenous Knowledge

The first influx of AI is also termed as the ‘expert’s system’ because it comprises algorithms and software which were developed by human experts, purely based on the knowledge that they possess. Their main attempt is to give these programs logical rules which can be decrypted. The AI system in the first wave is based on logical rules which make the system inspect the most important parameter and outreach the conclusion that will provide the most suitable action to solve a problem. These parameters are identified beforehand by human experts who make these systems deficient in deciphering new situations [10].

B. The Second Wave: The Measurable Learning

The second wave system consists of the least amount of human contribution. The statistical model is being created

by humans but these models ‘train’ themselves to produce high accuracy outcomes [11]. These models victoriously acknowledge the encompassing environment and perform predictive analysis which overshadows human intervention in various sectors. These models are sub-structures of artificial neural networks consisting of multiple layers that train themselves regarding the training dataset. The data is provided as the input and each layer process the data as well as the entire network and produces accurate results as output [12].

In closing, the second wave system produces outcomes that are better than the first wave systems. The interruption of these models in various fields spawns marvelous outcomes. These outcomes not only impetus for better decision-making and exclude human error but also save a lot of time, money and reduce human efforts. However, these models are ‘data-dependent’ as the data will determine the performance of the model; a poor training dataset will affect the learning and adapting ability according to the environment. Apart from that, these models lack the adroitness of explaining the internal functionality of the layer to the end-user [13].

C. The Third Wave: The Relevant Transformation

AI systems in third-wave themselves build models that will elucidate programs by unearthing logical rules which consolidate their decision-making. Third-wave frameworks are dependent on a few distinctive factual models to arrive at a more complete comprehension of a circumstance and also be able to explain the decision-making. Internal functionality of layers can be explained to the user which will overcome the problem of transparency in machine learning and deep learning [14].

Last but not least, this wave will boost AI, it will provide lucidity to the operator which makes the model more reliable but this wave is very sensitive to the quality of data, this not only affects the accuracy of models but also requires a reasonable amount of data for training.

TABLE I. PRESENT DEVELOPMENTS IN EACH WAVE

Waves	Developments
First wave	Good old fashioned AI(GOF AI)
Second wave	Deep learning and machine learning
Third-wave	XAI

There is also a fourth wave that tends to be self-governing. This wave has a very futuristic hypothesis in which machines will amalgamate everything and gain caliber to sight and retaliate according to the surroundings. This wave will empower AI to move and act gainfully, but it is still at the theoretical stage and requires a lot of work to be done to construct such types of models [15].

IV. THE FRAMEWORK OF XAI

XAI is an emerging range in AI and machine learning. The advancement of these technologies in various sectors require trust amongst humans regarding the decision made by these models. The models comprise heavy math and they are strenuous to decrypt which results in the reduction of explainability and peaking the complexity of models.

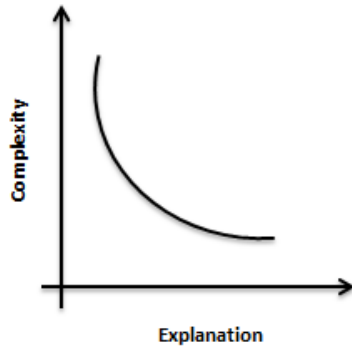


Fig.2. Representing the inversely proportional relationship between complexity and explanation.

The only way to gain certainty is by making the machine learning model pellucid. XAI systems are devices that create reports regarding how the model functions and attempt to clarify their working. There are numerous XAI frameworks accessible consisting of SHAP (Shapely Additive explanation), LIME (Local Interpretable Model-agnostic Explanation), ELI5, What-if-tool, AIX360, and Skaters amongst which SHAP and LIME are traditionally utilized. ELI5 is a python bundle that removes errors from machine learning classifiers, What-if-tool is matured by Google and helps in understanding the functioning of machine learning models, AIX360 and Skaters are open sources that empower model Interpretation for all types of model to assist one with building an Interpretable AI framework frequently required for true use-cases [16].

A. SHAP

SHAP can be used to elucidate several types of models including logistic regression, tree-based models, machine learning algorithms, and other complicated models including deep learning. SHAP is based on game theory's Shapley value and describes how output can be altered by using different features and connects credit allocations with local simplification by utilizing the help of Shapley's value and their connected augmentations [17].

B. LIME

LIME is akin to SHAP, but comparably, it is much faster in terms of computation and can explain black-box with more than one class. LIME's output is the mirror of the explanation, consisting of the contribution of features in the prediction of a data sample. It requires that the classifier carries out a capacity which takes in crude text and results in likelihood for each class. It can produce local justification for black-box models by producing locally bothered information and researching how models evolve toward this information. LIME is then ready to distinguish which specific annoyances were the most compelling in model predictions [18].

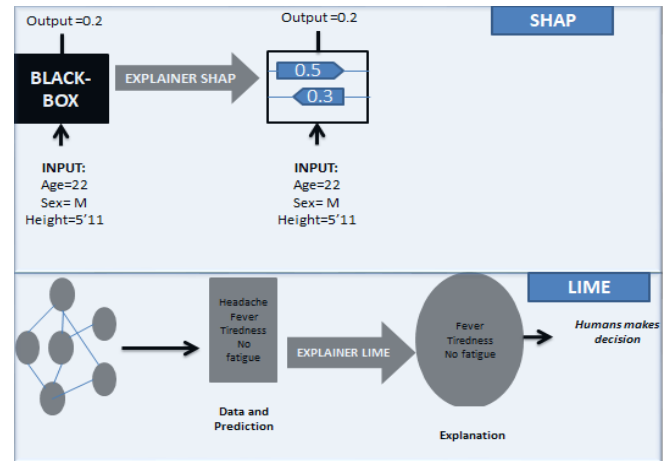


Fig.3. Working of SHAP and LIME

V. BLACK-BOX ALGORITHM

The development of technologies in various sector outcomes with a large amount of data that limits the human capability, which highlights the urgency of predictive and analysis tools so that better judgment and treatment options can be available at the same time, it'll also unfasten various prospects of research. Machine learning and deep learning algorithms, that are assigned with computer learning patterns for a specific dataset and produce results that are new and have unseen data [19].

Deep learning algorithms are an integral part of learning complex functions. These algorithms include representation-learning methods consisting of multiple-level representations procured by simple and non-linear modules that modify the representation at one level [12]. However, the current deep learning models have limitations that made the models deficient. The quality of data and accuracy of these models are directly proportional, huge data is required for training the model which consumes a lot of time and determines the prediction accuracy. Even though deep learning models are sovereign, these models are highly sensitive to errors, which makes them unreliable and biased [20].

The major reason which pulls back these models is the lack of transparency and explanation. Such types of models are termed as the 'black box' models which utilize processed data to train themselves that results in discreet relation between input and output. These models cannot provide the real reason for an output that raises issues regarding the internal functionality of layers [21]. In the conventional roadmap of such models, data is given as input and processed by computational layers which results in unseen data at output. Firstly, a training dataset is used to train the model with a particular learning process. Furthermore, the learning process expedites to learning function in which the input data is incorporated. After all the collective progress, the machine spawns the prediction at the output. The focal point reflects that prediction is being made but without any justification. Nevertheless, in the case of XAI, there is a new learning process that serves as the reason for the justification of the output. The additional layer empowers impartiality in decision-making. It not only assists to identify biases in the dataset but also features possible variables that can cause alteration in the final prediction.

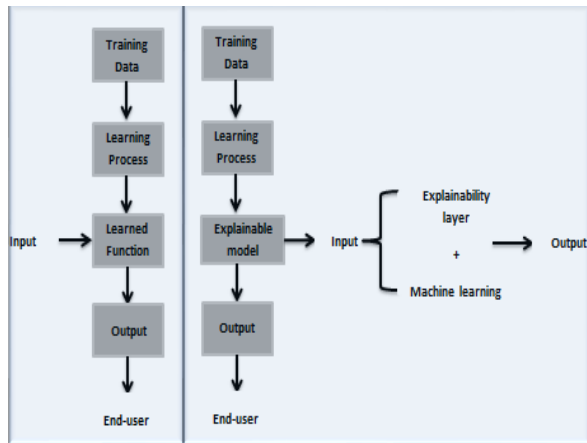


Fig.4. Conventional Machine learning model V/S XAI model

The XAI is considered as the region of convergence of various fields. One of the regions is end-user clarification with social science that provides XAI to gain discernment abilities. Another region is the human-machine interface, where it can show the capacity to clarify; because logical computerized reasoning requires an extremely significant level of collaboration with the user [22].

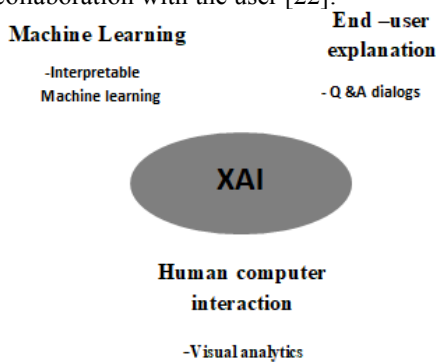


Fig.5. Convergence of XAI with various fields

VI. CHALLENGES FACED BY XAI

Issues with Biases Data: The most conventional wrangle in the case of impartial machine learning is to acknowledge the fact that poor quality data will produce poor quality outcomes. The procedure of generating, collecting, and processing the data increase the chances of human bias which will lead to biased and unreliable outcomes. Selection bias will restrict the data and inhibits the ability to be illustrative and miscellaneous; inherent and measurement bias puts a limit on the precise data; confirmation biased makes the inaccuracies of data opaque to the auditor [23]. The algorithm of machine learning is the outcome of their data; data that frequently act as a mirror of the society. This results in an unavoidable societal bias which creates data that reflects bigotry and enhances stereotypes. The information assortment stage is a long way from impartiality. There are biases connected to the properties that are settled on, the objective variable that is chosen, and the examples that are picked to be incorporated.

The variability in the sample size is considered a major issue, this arises unmanageable plight in which inadequate data results in poor models having inaccurate outcomes.

Moreover, the data collectors hold the veto to determine the configuration of the data set. The life and professional preferences of data collectors determine which attribute dominates while contemplating the target variable. This approach is not germane to others with different frameworks.

The final stage includes designing of machine learning model with processed data. This stage is very sensitive to human bias, steps like feature engineering can affect the final algorithm of machine learning. The management of imperfect or missing data can radically alter the outcome of the algorithm [24].

The raw and processed data can enhance the ability of XAI and allow it to divulge the disparity that may exist. Considering this, XAI can serve in these four areas:

- XAI can sever in identifying imbalance that relates to oversampling
- XAI can classify attributes that have an impact on both local and global judgment.
- Processing issues can be highlighted with XAI, which can alter the final decision-making.
- The XAI apparatuses can consider the effect of client-marked sensitive characteristics on the model presentation.

Issues with options and designing the machine learning:

The biases are one of the factors that influence the conclusion of the algorithm; others include the selection of machine learning models and the limiting justification that fortuitously introduce components results in affecting the outcome. The selection of machine learning affects the quality of outcomes. For instance, the link between the variables is not visible in regression models. The individual variable is appraised liberally to the other available variables. In the other cases in which the selection of biases occurs within the data, models like Bayesian classifiers are much recommended to handle this kind of data. The selection of methods can be inaccurate. Even more generally, the supervised method is put forward in labeled data; nevertheless, it is more suitable to use an unsupervised method before or in place of the supervised method [25]. Designing machine learning includes functions that hone the algorithm. These functions can unintendedly raise issues regarding fairness in the model. In addition, it depends on the situation to choose the selection of evaluation and functionality optimization in regards to the fairness of the model. It is significant while considering reasonableness that user can choose their assessment metric.

The fairness issues raise a couple of expectations from XAI:

- The XAI can identify the effect caused by the selection of machine learning and optimization which struck the overall performance.
- The XAI apparatuses think about some measurement of decency in assessing the global presentation of the subsequent calculation.

Issues with the presentation of XAI tools: The potential of XAI is calculated on the level of their explanation. The background of the user should not interfere while employing the XAI tool. The professional operates XAI tools to get the perception of global and local practices of their model. Into

the bargain, the XAI tools should only be built for the audience they are intended for. AI specialists, partners, and shoppers all use the innovation concerning their assumptions and what to conclude about intelligibility and impartiality [24].

VII. APPLICATIONS

Financial Forecasting: The foremost step in machine learning manufacturing is feature extraction and plays a pivotal role in the identification and supports the model in classification tasks. In the financial domain, this step is important and more complex to extract features from financial data. With the introduction of XAI, different feature selection strategies can be created in an applied monetary setting where we need to anticipate the following day returns for a bunch of input stocks [26].

Healthcare: The introduction of AI in the health sector always contributed to better decision-making and providing various treatment options. The merging of AI and wearable devices (E.g., Fitbit) has successfully predicted the health condition through analyzing the health data detained by the device. This combination has very robust applications in the field of healthcare but lacks when it comes to operating the black-box decision created by AI. and raises questions of trust. The introduction of XAI and its approaches unfold the explanation behind the decision of AI. XAI analyzes the health data by AI system and focuses on the path to attain the responsibility, lucidity, tracking results to make these models more reliable and trustworthy[27].

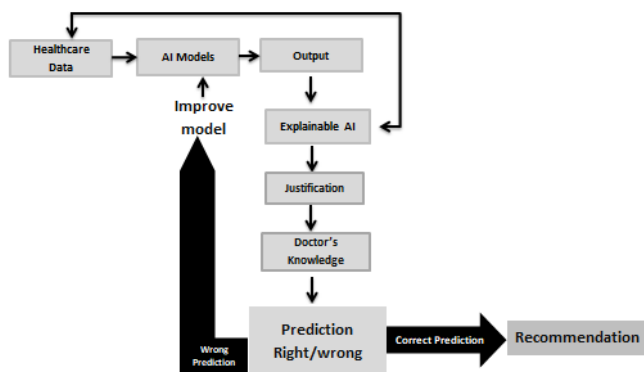


Fig.6. XAI model in healthcare

Industry: The incorporation of AI in different sectors has led to bias caused by the models. These models raise questions regarding transparency and accountability. In the case of industrial application, the analysis becomes economically sensitive which highlights the need for explainability. The XAI assists them with emphasizing trust and comprehending models at scale [28].

Data Quality: The role of explainability in Data Quality with regards to data-based machine learning helps in providing internal functionality. XAI not only highlights the model but also examines the data that construct the models. Through explainability, there is always a margin in defining, identifying, and elucidating errors in data before constructing the model and also provides the repair action.

Overall, explainability can be utilized as a reason for characterizing information with regards to machine learning and estimating or measuring them [29].

VIII. CONCLUSION

Today, every domain of life whether it's healthcare, finance, or industry, is driven by data; data that is unstructured and hard to manage. AI and its models come with a guarantee to analyze the data at very high accuracy that intensifies these domains. The role of AI in performing critical tasks and providing options for better judgment has increased rapidly and possession of humans over these models has been swelled up. However, these models are not able to explain the reason behind their output to the end-user. This not only raises the trust question but also questions the fairness of the model. Therefore, it is very necessary to make these models more reliable and crystalline. With the introduction of XAI, it becomes possible for humans to decipher the reason behind the outcome.

The repercussions of AI-founded output are directly proportional to the demand for XAI. The features that can transmute the final product can be highlighted by XAI that provides a fruitful knowledge of models to the operator through which the future use of such model can be determined further. XAI is based on several frameworks amongst which SHAP and LIME are conventionally used. But, the advancement of any technology is commensurate to its counter challenges. The challenges faced by XAI include issues with prejudiced data which reflects discrimination in the output and there are issues regarding creating algorithms for a machine learning model which cause inexorable creative bias in the model. Recent research highlight the role of XAI in different sections of the actual world, but it also suggests that it requires a lot of work which is needed to be done for utilizing the maximum capacity of such tools. The headway in innovations and disadvantages are straightforwardly corresponding to each other. Hence, it is exceptionally fundamental to recognize the drawback as well as to unsheathe the perks.

Technology limits human potential but also serves humans, today the AI models use unexpressed mastery which makes these models blurry to humans, in response technology has come up with XAI that uses explicit knowledge for humans. The environment of AI has been divided into explicit and implicit knowledge. Analysts are now bringing these two into a harmony that could act as catalysts for future grand challenges. It also features the desperation of technophiles who can separate the maximum capacity of XAI that furnishes different domains and future research.

REFERENCES

- [1] K. Alikhademi, B. Richardson, E. Drobina, and J. E. Gilbert, "Can Explainable AI Explain Unfairness? A Framework for Evaluating Explainable AI," Jun. 2021, Accessed: Dec. 28, 2021. [Online]. Available: <http://arxiv.org/abs/2106.07483>
- [2] R. K. E. Bellamy *et al.*, "AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias," Oct. 2018, Accessed: Dec. 28, 2021. [Online]. Available: <http://arxiv.org/abs/1810.01943>
- [3] J. Dodge, Q. Vera Liao, Y. Zhang, R. K. E. Bellamy, and C. Dugan, "Explaining models: An empirical study of how explanations impact fairness judgment," *International Conference on Intelligent User Interfaces, Proceedings IUI*, vol. Part F147615, pp. 275–285, 2019.

- [4] Wojciech Samek, Thomas Wiegand, and Klaus-Robert Müller, "Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models," *arXiv preprint arXiv:1708.08296* (2017), 2017.
- [5] N. Haefner, J. Wincent, V. Parida, and O. Gassmann, "Artificial intelligence and innovation management: A review, framework, and research agenda," *Technological Forecasting and Social Change*, vol. 162, p. 120392, Jan. 2021, doi: 10.1016/J.TECHFORE.2020.120392.
- [6] Shane T. Mueller and Elizabeth S. Veinott, "Principles of Explanation in Human-AI Systems," 2021.
- [7] Yann LeCun and Yoshua Bengio, "Deep learning," May 2015.
- [8] Mervin Jeff, "When will we get there?," *Science V*, 2017.
- [9] Jenkins C and Lopresti D, "Next Wave Artificial Intelligence: Robust, Explainable, Adaptable, Ethical, and Accountable," 2020.
- [10] Kristian Kersting, "Rethinking Computer Science Through AI," Dec. 2020.
- [11] Kai-Fu Lee, "The Four Waves of A.I.," Nov. 23, 2018.
- [12] Y. Lecun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, 2015.
- [13] Feiyu Xu and Hans Uszkoreit, "Explainable AI: A Brief Survey on History, Research Areas, Approaches and Challenges," Sep. 2019.
- [14] Holger Hoos and Kristian Kersting, "The third wave of artificial intelligence," Dec. 2020.
- [15] "Artificial Intelligence Tech Will Arrive in Three Waves." <https://futurism.com/artificial-intelligence-tech-will-arrive-in-three-waves> (accessed Dec. 28, 2021).
- [16] Danding Wang and Qian Yang, "Designing Theory-Driven User-Centric Explainable AI," *CHI '19: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, May 2019.
- [17] M. Chromik, "reSHAPe: A Framework for Interactive Explanations in XAI Based on SHAP," *European Society for Socially Embedded Technologies (EUSSET)*, 2020.
- [18] J. Dieber, "Why model why? Assessing the strengths and limitations of LIME," Nov. 2020.
- [19] Došilović FK, "Explainable artificial intelligence: A survey.," *41st International convention on information and communication technology, electronics and microelectronics (MIPRO). IEEE*, 2018.
- [20] Schnack H, "interpretability in machine learning: From measurements to features. In: Machine Learning.," *Academic Press*, 2020.
- [21] Samek W and Wiegand T, "Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models," 2017.
- [22] E. Dağlarlı, "Explainable Artificial Intelligence (xAI) Approaches and Deep Meta-Learning Models," *Advances and Applications in Deep Learning*, Jun. 2020, doi: 10.5772/INTECHOPEN.92172.
- [23] J. Dodge, Q. Vera Liao, Y. Zhang, R. K. E. Bellamy, and C. Dugan, "Explaining models: An empirical study of how explanations impact fairness judgment," *International Conference on Intelligent User Interfaces, Proceedings IUI*, vol. Part F147615, pp. 275–285, 2019.
- [24] K. Alikhademi, B. Richardson, E. Drobina, and J. E. Gilbert, "Can Explainable AI Explain Unfairness? A Framework for Evaluating Explainable AI," Jun. 2021, Accessed: Dec. 29, 2021. [Online]. Available: <http://arxiv.org/abs/2106.07483>
- [25] E. Toreini, M. Aitken, K. Coopamootoo, K. Elliott, C. G. Zelaya, and A. van Moorsel, "The relationship between trust in AI and trustworthy machine learning technologies," *FAT* 2020 - Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 272–283, Jan. 2020.
- [26] Salvatore Carta, "Explainable AI for Financial Forecasting".
- [27] U. Pawar, D. O'Shea, S. Rea, and R. O'Reilly, "Explainable AI in Healthcare," *2020 International Conference on Cyber Situational Awareness, Data Analytics and Assessment, Cyber SA 2020*, Jun. 2020.
- [28] Krishna Gade and Sahin Cem Geyik, "Explainable AI in Industry," *the 25th ACM SIGKDD International Conference*, 2019.
- [29] L. Bertossi and F. Geerts, "Data Quality and Explainable AI," *Journal of Data and Information Quality*, vol. 12, no. 2, May 2020.

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

CANDIDATE'S DECLARATION

I Nakul Tanwar, Roll Number: 2K20/MSCBIO/40, student of M.Sc. Biotechnology, hereby declare that the work which is presented in the Major Project entitled —Explaining developmental neurotoxicity by XAI 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: Explainable AI; Are we there yet?

Author Names: Tanwar, Nakul, and Hasija, Yasha

Name of Conference: 2022 IEEE Delhi Section Conference (DELCON)

Conference Date and Venue: 11-13 Feb 2022 at Netaji Subhas University of Technology

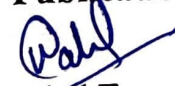
Registration: Done

Status of Paper: Published

Date of Paper Communication: 31th December 2021

Date of Paper Acceptance: 16th January 2022

Date of Paper Publication: 21st April 2022

Date: 6/5/22 
Nakul Tanwar

DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College Of Engineering)
Bhawana Road, Delhi-110042

Certificate

I hereby certify that the Project Dissertation titled "**Explaining Developmental Neurotoxicity By XAI**" which is submitted by **Nakul Tanwar (2K20/MSCBIO/40)**, Department of Biotechnology, Delhi Technological University, Delhi in partial fulfillment of the requirement for the award of the degree of Master of Science is recorded 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 any diploma to this university or elsewhere.

Place: Delhi

Date:

Yasha Hasija
05/05/22

Prof. Yasha Hasija
(Supervisor)
Professor

Department of Biotechnology
Delhi Technological University

Pravir Kumar
05/05/2022

Prof. Pravir Kumar
Head of Department

Department of Biotechnology
Delhi Technological University

Acknowledgement

I would like to express my gratitude to 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 been 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 she has offered me. Her insightful feedback pushed me to sharpen my thinking and brought my work to a higher level.

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

I would like to thank MR. Rajkumar Chakraborty without whom I would not have been able to complete this research, and without whom I would not have made it through my master's degree. I would like to thank the following people for helping with this research project: Miss. Neha Kumari, Mrs. Jaishree Meena, and Mrs. Priya

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



Nakul Tanwar