शोध पत्र / Research Paper

मशीन लर्निंग का उपयोग करके प्रारंभिक चरण मधुमेह जोखिम भविष्यवाणी

Early Stage Diabetes Risk Prediction using Machine Learning

पार्थ घोष

Partha Ghosh

Dept. of Computer Sc. and Engineering
Govt. College of Engineering and Ceramic Technology, Kolkata, India
parth_ghos@redffimail.com

सारांश

दुनिया भर में 422 मिलियन से अधिक लोगों को प्रभावित करने वाली सबसे पुरानी पुरानी जानलेवा बीमारी मधुमेह है। टाइप 2 मधुमेह मुख्य रूप से जीवन शैली के फैसले और पर्यावरणीय कारकों के कारण होता है। यह एक धीमी गति से बढ़ने वाली बीमारी है जो एक बीमारी में विकसित होने से बहुत पहले ही चयापचय कारकों को विकसित करना शुरू कर देती है और एक उपवास चीनी परीक्षण द्वारा औपचारिक रूप से निदान किया जाता है। डेटासेट [dataset] में 520 रोगियों के मधुमेह से संबंधित संकेतों के रिकॉर्ड हैं जिनका उपयोग किया गया था। इसमें लोगों के डेटा शामिल हैं, जैसे कि उम्र, लिंग और लक्षण जो मधुमेह का कारण हो सकते हैं। मैंने नैवे बेस क्लासिफायर (एनबी), लॉजिस्टिक रिग्रेशन क्लासिफायर (एलआर), जे 48 एल्गोरिथम, रैंडम फॉरेस्ट (आरएफ) और मल्टी-लेयर परसेप्ट्रॉन (एमएलपी) एल्गोरिदम का उपयोग करके डेटासेट का विश्लेषण किया। और एमएलपी को इस डेटासेट पर दस गुना क्रॉस-वैलिडेशन और प्रतिशत स्प्लिट मूल्यांकन तकनीकों को लागू करने के बाद सबसे अच्छा सटीकता पाया गया था। MLP इस डेटासेट [dataset] और बहुत कम संख्या में मिसकैरेज के साथ 98% सटीकता प्राप्त करता है।

मेडिकल डोमेन के मामले में फजी और अनिश्चित डेटा को संभालना एक और महत्वपूर्ण मुद्दा है। हाल के वर्षों में चिकित्सा डेटा में अनिश्चितता से निपटने पर अधिक ध्यान दिया गया है। इस कार्य में मधुमेह के निदान के लिए अनुकूली तंत्रिका-फजी अनुमान प्रणाली (ANFIS) [10] का उपयोग किया गया है। यह अस्पष्टता (फर्जी) में अनिश्चितता को मॉडल करने के लिए तंत्रिका नेटवर्क के साथ अस्पष्ट (फजी) तर्क की सीखने की क्षमता को एकीकृत करता है। फ़ज़ी लॉजिक का उपयोग अनिश्चित परिदृश्यों को मॉडल करने के लिए किया जाता है और वह मॉडल तंत्रिका नेटवर्क द्वारा सीखा जाता है। एएनएफआईएस में तंत्रिका नेटवर्क को ताकागी-सुजेनो फजी इंट्रस सिस्टम के साथ एकीकृत किया गया है और यह गणितीय गणनाओं पर आधारित है जो जटिल समस्याओं को हल कर सकता है। पिमा इंडियन डायबिटिक डेटासेट (PIDD) [dataset] को MATLAB में वर्गीकरण के लिए प्रशिक्षित और परीक्षण किया गया था। यहां, मैंने मधुमेह के निदान के लिए इस पद्धति का उपयोग इसकी मजबूत अनिश्चितता से निपटने की क्षमता और व्याख्यात्मकता का उपयोग करके किया है ताकि आशाजनक वर्गीकरण प्रदर्शन प्राप्त किया जा सके।

Abstract

The fastest chronic life-threatening disease ffaecting more than 422 million people globally is diabetes. Type 2 diabetes is mainly due to lifestyle decisions and environmental factors. It is a slow–growing disease which starts to develop metabolic factors long before they evolve into a disease and is formally diagnosed by a fasting sugar test. There are records of indications related to diabetes from 520 patients in the dataset [dataset] that was used. It contains data on people, such as age, sex and symptoms that may lead to diabetes. I analyzed the dataset using Naive

Bayes Classifier (NB), Logistic Regression Classifier (LR), J48 Algorithm, Random Forest (RF) and Multi–Layer Perceptron (MLP) Algorithm. And MLP was found to have the best accuracy on this dataset after applying tenfold Cross–Validation and Percentage Split evaluation techniques. MLP achieve 98% accuracy with this dataset [dataset] and very few numbers of misclassification.

In the medical field, dealing with ambiguous and uncertain data is also a significant concern. In recent years, there has been a greater focus on handling uncertainty in medical data. In this study, the adaptive neuro–fuzZy inference system (ANFIS) [10] was employed to diagnose diabetes. It combined fuzZy logic's learning capabilities with neural networks to describe uncertainty in expressiveness. To represent uncertain circumstances, fuzZy logic is used, and the model is trained by a neural network. The neural network of ANFIS is based on mathematical computations and is linked with the Takagi–Sugeno fuzZy inference system to tackle complicated problems. MATLAB was used to train and test the Pima Indian Diabetic Dataset (PIDD) [dataset] for classification. I've utilised this method to diagnose diabetes by leveraging its great uncertainty–handling capabilities and interpretability to produce good classification results.]

कीवर्ड : मधुमेह का खतरा; Naive Bayes Classifier; मल्टी-लेयर पर्सेप्ट्रॉन (एमएलपी); बेतरतीब जंगलय लॉजिस्टिक रिग्रेशन क्लासिफायरियर; लक्षण; अनिश्चितता।

Keywords: Diabetes risk; Naive Bayes Classifier; Multi–Layer Perceptron (MLP); Random forest; Logistic Regression Classifier; Symptom; Uncertainty.

1. परिचय

मधुमेह निदान को मात्रात्मक अनुसंधान के लिए एक कठिन मुद्दा माना जाता है। कुछ सीमाओं के कारण, कुछ मापदंडों जैसे A1c [7], हैक्टिकली, वाइट ब्लड सेल काउंट, फैटी एसिड ऑक्सीडेशन और हेमेटोलॉजिकल इंडेक्स [8] को अपर्याप्त दिखाया गया है। इलेक्ट्रोफोरो. सिस द्वारा भविष्यवाणी किए जाने पर विटामिन सी का सेवनA1c बढ़ा सकता है, लेकिन जब क्रोमैटोग्राफी द्वारा भविष्यवाणी की जाती है, तो मान कम हो सकता है। अधिकांश अध्ययनों ने संकेत दिया है कि उच्च रक्तचाप के दौरान भड़काऊ प्रतिक्रिया सफेद रक्त कोशिकाओं की अधिक संख्या का कारण बनती है [9]A मधुमेह की सटीक भविष्यवाणी करने के लिए, एक एकल पैरामीटर सिर्फ बहुत प्रभावी नहीं है और निर्णय लेने के अभ्यास में भ्रमित हो सकता है। प्रारंभिक बिंदु पर मधुमेह की कुशलता से भविष्यवाणी करने के लिए, विभिन्न मापदंडों को समेकित करने की आवश्यकता है। हमारे विश्लेषण में, महत्वपूर्ण विशेषताओं और विभिन्न विशेषताओं के सहयोग से मधुमेह की भविष्यवाणी की जाती है।

मशीन लर्निंग एल्गोरिदम जो बेहतर निर्णय लेने की प्रक्रिया के लिए लाभकारी जानकारी प्रकट करने के लिए अनिश्चितता को मॉडल कर सकते हैं, बहुत काम का होगा। आम तौर पर, अनिश्चितता दो कारणों से हो सकती है : डेटा (शोर) अनिश्चितता और मॉडल अनिश्चितता (जिसे महामारी संबंधी अनिश्चितता भी कहा जाता है)। माप की अशुद्धि के कारण लेबलों के बीच शोर होने की संभावना है, जिससे एलिएटोरिक अनिश्चितता हो सकती है। इस बीच, मॉडल अनिश्चितता को दो मुख्य प्रकारों में विभाजित किया जा सकता हैरू संरचना अनिश्चितता और मॉडल मापदंडों में अनिश्चितता। संरचनात्मक अनिश्चितता में, हम उपयोग की जाने वाली मॉडल संरचना के प्रकार का पता लगाते हैं और हमारे प्रस्तावित मॉडल को एक्सट्रपोलिंग और/या इंटरपोलिंग के लिए निर्दिष्ट करते हैं। दूसरे प्रकार में, यानी मॉडल मापदंडों में अनिश्चितता, अधिक सटीक भविष्यवाणियों के लिए इष्टतम मॉडल मापदंडों का चयन किया जाता है। चिकित्सा विज्ञान में अनिश्चितताओं को संभालने के लिए सबसे आम एल्गोरिदम हैं बायेसियन इंट्रस, फजी सिस्टम्स, मोंटे कार्लो सिमुलेशन, रफ क्लासिफिकेशन, डेम्पस्टर-शेफर थ्योरी और इम्प्रेसिस प्रोबेबिलिटी। इसलिए, शोधकर्ताओं ने निदान के लिए कई तरीके प्रस्तावित किए हैं, जिनमें से

एक अस्पष्ट (फजी) आधारित विशेषज्ञ प्रणाली [11] 12, है जो अस्पष्ट (फजी) नैदानिक डेटा पर विचार करते हुए चिकित्सा क्षेत्र में निदान के लिए उपयुक्त है। हालांकि, इसका एक अधिक शक्तिशाली विस्तार है जिसे अनुकूली न्यूरो-फजी इंट्रस सिस्टम (ANFIS) [10] कहा जाता है, जो अस्पष्टता में अनिश्चितता को मॉडल करने के लिए तंत्रिका नेटवर्क के साथ फजी लॉजिक की सीखने की क्षमता को एकीकृत करता है। फ़ज़ी लॉजिक का उपयोग अनिश्चित परिदृश्यों को मॉडल करने के लिए किया जाता है और वह मॉडल तंत्रिका नेटवर्क द्वारा सीखा जाता है।

फ़ज़ी सिस्टम के साथ एकीकृत तंत्रिका नेटवर्क के माध्यम से अनुकूलन क्षमता कारक पर विचार करके वर्गीकरण सटीकता में सुधार करने के लिए, इस शोध पत्र में अनुकूली न्यूरो फ़ज़ी इंफ़ेक्शन सिस्टम (ANFIS) का उपयोग किया जाता है। इस प्रणाली का उपयोग करने के कुछ फायदे हैं। यह एक जटिल प्रणाली के व्यवहार को चित्रित करने के लिए मानव विशेषज्ञता के बिना फजी अगर-फिर नियमों को बढ़ाता है। तेजी से अभिसरण समय। यह सदस्यता कार्यों का उपयोग करता है। तेजी से सीखने की क्षमता और एक प्रक्रिया की गैर-रेखीय संरचना को पकड़ने की क्षमता।

फ़ज़ी आधारित मॉडल ऐसी स्वास्थ्य देखभाल प्रणा. लियों के लिए लाभकारी लागत प्रभावी प्रणालियों में से एक हैं। यह शक्तिशाली तर्क क्षमताओं के साथ इन अनिश्चितताओं को दूर करके सटीक समाधान प्रदान करता है। एडेप्टिव न्यूरो-फजी इंट्रस सिस्टम (ANFIS) फजी लॉजिक और न्यूरल नेटवर्क सिद्धांतों को एक फ्रेम में इंटरपोलेशन और सीखने की क्षमता दोनों के लाभ के साथ एकीकृत करता है। नतीजतन, इन संयुक्त सुविधाओं के साथ गैर-रेखीय कार्यों का कुशलतापूर्वक अनुमान लगाया जाता है। एएनएफआईएस में तंत्रिका नेटवर्क को ताकागी-सुजेनो फजी इंट्रस सिस्टम के साथ एकीकृत किया गया है और यह गणितीय गणनाओं पर आधारित है जो जटिल समस्याओं को हल कर सकता है।

ज्ञान और निदान डेटा के साथ इन विशेषज्ञ प्रणालियों को लागू करना न्यूनतम त्रुटि के साथ कुशल निदान परिणाम प्रदान कर सकता है।

2. संबंधित कार्य

शेट्टी एट अल। [1] नियोजित ज्ञछछ और मधुमेह भविष्यवाणी के लिए Naïve Bayes की रणनीति का उपयोग किया गया है। एक विशेषज्ञ सॉफ्टवेयर अनुप्रयोग के रूप में, उनकी कार्यप्रणाली को लागू किया गया था, जहां उपयोगकर्ता चिकित्सा रिकॉर्ड के संदर्भ में प्रतिक्रिया प्रदान करते हैं और यह निष्कर्ष निकालते हैं कि रोगी मधुमेह है या नहीं। सिंह एट अल द्वारा विभिन्न प्रकार के डेटासेट पर विभिन्न एल्गोरिदम लागू किए गए थे। [2] वे KNN, रैंडम फ़ॉरेस्ट और Naesve Bayesian से एल्गोरिदम का उपयोग करते थे। मूल्यांकन के लिए, के-गुना क्रॉस-सत्यापन विधि का उपयोग किया गया था। मधुमेह के निदान के लिए, अहमद ने रोगी डेटा और उपचार योजना आयामों का उपयोग किया। नाओवे बेय्स, लॉजिस्टिक और जे 48 इन तीन एल्गोरिदम को विधियों में जोड़ा गया था। एंटनी एट अल।[3] मधुमेह की भविष्यवाणी के लिए चिकित्सा साक्ष्य का उपयोग किया। डेटा के पूर्व-प्रसंस्करण के बाद, Naïve Bayes, फ़ंक्शन-आधारित बहुपरत अवधारणात्मक (MLP), और निर्णय-ट्री-आधारित यादृच्छिक वन (RF) प्रक्रियाएँ लागू की गईं। अतिरिक्त विशेषताओं को बाहर करने के लिए, सहसंबंध आधारित सुविधा चयन दृष्टिकोण का उपयोग किया गया था। सीखने का एक मॉडल तब भविष्यवाणी करता था कि रोगी मधुमेह था या नहीं। अन्य मशीन लर्निंग एल्गोरिदम की तुलना में, जब Naïve Bayes एल्गोरिथ्म एक पूर्व-प्रसंस्करण प्रणाली के माध्यम से उपयोग किया जाता है तो परिणाम बढ़ाए गए थे। प्रारंभिक मधुमेह भविष्यवाणी के लिए पीआईडी डेटासेट का विस्तार करके, अमीना एट अल [4] ने विभिन्न डेटा खनन एल्गोरिदम की तुलना की। तल्हा एट अल [5] उनके परिणाम मधुमेह और बॉडी मास इंडेक्स (बीएमआ. ई) और ए-प्राथमिकता प्रक्रिया के माध्यम से निकाले गए ग्लूकोज की मात्रा के बीच एक स्पष्ट संबंध का सुझाव देते हैं। मधुमेह की भविष्यवाणी के लिए, कृत्रिम तंत्रिका नेटवर्क (एएनएन), यादृच्छिक वन (आरएफ) और के-साधन क्लस्टरिंग तकनीक पेश किए गए थे।।छछ विधि ने 75.7% की अधिकतम सटीकता दी। इस्लाम

एट अल। ख्6, दिखाया गया कि रैंडम फॉरेस्ट रेजिस्टर सही है और 97.4% सटीकता हासिल की जा सकती है। वे निर्णय पेड़ों और अन्य शिक्षार्थियों के साथ इसकी तुलना करते हैं। दोनों 10-गुना क्रॉस-मान्यता और 80/20 विभाजन का उपयोग किया जाता है।

इसके अलावा, कुछ शोधकर्ताओं ने न्यूट्रोसोफिक लॉजिक [13] न्यूट्रोसोफिक ग्राफ, हाइपरग्राफ, और अन्य सॉफ्ट कंप्यूटिंग तकनीकों के गुणों का उपयोग करके चिकित्सा निदान डेटा सेट में अनिश्चितता और अस्पष्टता को मापने का प्रयास किया। सिंह [14] द्वारा विस्तृत थ्री-वे फ़ज़ी कॉन्सेप्ट जाली के कैलकुलस का उपयोग करते हुए मेडिकल डायग्नोसिस डेटा सेट का पर्याप्त विश्लेषण।

3. प्रायोगिक परिणाम और चर्चा

3ण्1. डेटासैट का इस्तेमाल किया

डेटासेट एक महत्वपूर्ण प्रश्नावली से संकलित किया गया था और एक डॉक्टर की देखरेख में पूरा किया गया था। इस अध्ययन के लिए अन्य मधुमेह संबंधी कारकों पर विचार किया गया था ताकि पूर्व-मधुमेह व्यक्तियों का निदान किया जा सके। मधुमेह चिकित्सा डेटासेट को कैलिफोर्निया विश्वविद्यालय, इरविन (यूसीआई) मशीन लर्निंग रिपॉजिटरी [dataset] से एकत्र किया गया था। इसके अलावा पीमा इंडियन डायबिटिक डेटासेट (PIDD) का भी इस्तेमाल किया गया था [dataset]।

3.2. डेटा की व्याख्या

डेटासेट में कोई गुम मान नहीं था। इस डेटासेट के डेटा फॉर्म में 16 विशेषताएँ हैं जिनका उपयोग परिणामों, कक्षा चर + ve की भविष्यवाणी करने के लिए किया गया था, जिसका अर्थ है कि एक व्यक्ति मधुमेह और वर्ग चर -ve है, जिसका अर्थ है कि एक रोगी में कोई मधुमेह नहीं है।

हमारे पास मधुमेह की कक्षा की भविष्यवाणी करने के लिए उपयोग की जाने वाली 16 विशेषताएं हैं जैसा कि तालिका-1 में दिखाया गया है। उम्र के अपवाद के साथ सभी विशेषताओं में दो अलग-अलग परिणामों के साथ स्पष्ट डेटा है। रोगियों की आयु 16 से 90 वर्ष के बीच है।

Attributes	Description
Age	16-90
Sex	1. Male, 2.Female
Polyuria	1.Yes, 2.No.
Polydipsia	1.Yes, 2.No.
sudden weight loss	1.Yes, 2.No.
weakness	1.Yes, 2.No.
Polyphagia	1.Yes, 2.No.
Genital thrush	1.Yes, 2.No.
Visual blurring	1.Yes, 2.No.
Itching	1.Yes, 2.No.
Irritability	1.Yes, 2.No.
Delayed healing	1.Yes, 2.No.
Partial paresis	1.Yes, 2.No.
Muscle stiffness	1.Yes, 2.No.
Alopecia	1.Yes, 2.No.
Obesity	1.Yes, 2.No.
Class	1.Positive 2.Negative

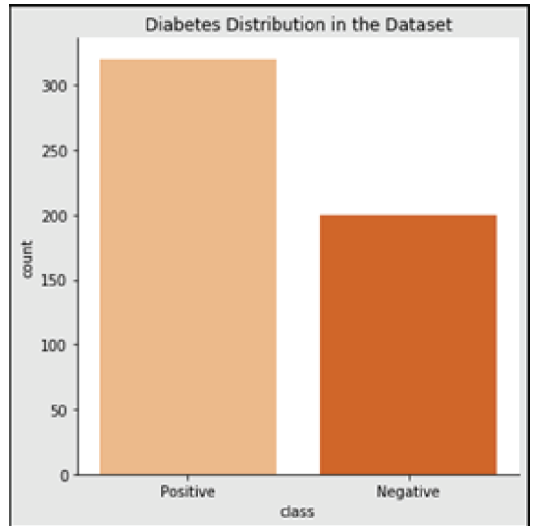
तालिका-1 : 16-सुविधाएँ या विशेषताएँ
[Table-1 : 16- features or attributes]

	Age	Gender	Polyuria	Polydipsia	sudden weight loss	weakness	Polyphagia	Genital thrush	visual blurring	Itching	Irritability	delayed healing	partial paresis	muscle stiffness	Alopecia	Obesity	class
0	40	Male	No	Yes	No	Yes	No	No	No	Yes	No	Yes	No	Yes	Yes	Yes	Positive
1	58	Male	No	No	No	Yes	No	No	Yes	No	No	No	Yes	No	Yes	No	Positive
2	41	Male	Yes	No	No	Yes	Yes	No	No	Yes	No	Yes	No	Yes	Yes	No	Positive
3	45	Male	No	No	Yes	Yes	Yes	Yes	No	Yes	No	Yes	No	No	No	No	Positive
4	60	Male	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Positive

तालिका-2 : पहले पांच पंक्तियों को प्रदर्शित कर
[Table-2 : display first five rows]

तालिका-2 से ऊपर की पहली पाँच पंक्तियाँ प्रदर्शित होती हैं। डेटा एक सुव्यवस्थित प्रारूप में होता है, जिसमें प्रत्येक पंक्ति में एक अवलोकन का उपयोग करके, कॉलम में वैरिएबल मान होते हैं। सुविधा का आकार (520, 17) है। 38.5 प्रतिशत रोगियों को मधुमेह नहीं था और 61.5 प्रतिशत को मधुमेह था जैसा कि चित्र-1 में दिखाया गया है।

37% और 63% रोगी क्रमशः महिला और पुरुष थे। 90% महिलाओं को मधुमेह था जबकि 45% पुरुषों को मधुमेह था।



चित्र-1 : नमूने में मधुमेह के वितरण के लिए दृश्य

[Figure-1 : Visualization for distribution of diabetes in the mple]

3.3. डेटा में हेरफेर

मशीन सीखने और सहसंबंध कार्यों के लिए कार्यों को करने के लिए, डेटासेट को गैर-संख्यात्मक लेबल से संख्यात्मक आइटम में परिवर्तित किया गया। इसलिए पहले हमने संख्यात्मक डेटा (आयु) को हटा दिया है और शेष श्रेणीब) डेटा को तालिका -3 में दिखाए गए संख्यात्मक डेटा में बदल दिया है।

	Gender	Polyuria	Polydipsia	sudden weight loss	weakness	Polyphagia	Genital thrush	visual blurring	Itching	Irritability	delayed healing	partial paresis	muscle stiffness	Alopecia	Obesity	class
0	1	0	1	0	1	0	0	0	1	0	1	0	1	1	1	1
1	1	0	0	0	1	0	0	1	0	0	0	1	0	1	0	1
2	1	1	0	0	1	1	0	0	1	0	1	0	1	1	0	1
3	1	0	0	1	1	1	1	0	1	0	1	0	0	0	0	1
4	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1

तालिका-3 : हटाए गए संख्यात्मक डेटा ('आयु') और डेटासेट को संख्यात्मक वस्तुओं में परिवर्तित किया गया।

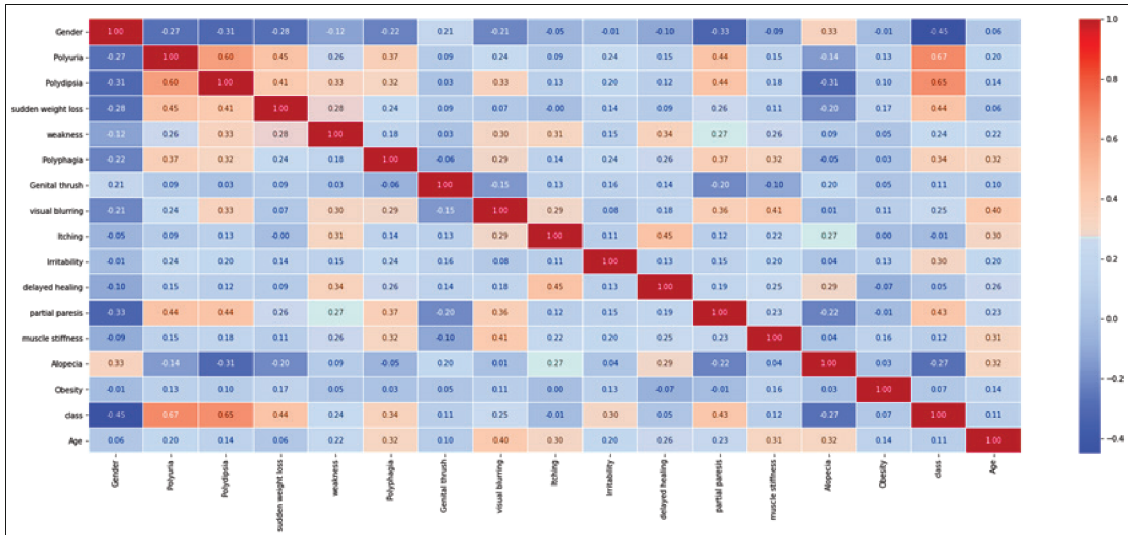
[Table-3: Deleted numerical data ('Age') and dataset was converted into numeric items.]

फिर से हमने ष्आयुष् को डेटासेट में वापस जोड़ा जैसा कि तालिका -4 में दिखाया गया है।

	Gender	Polyuria	Polydipsia	sudden weight loss	weakness	Polyphagia	Genital thrush	visual blurring	Itching	Irritability	delayed healing	partial paresis	muscle stiffness	Alopecia	Obesity	class	Age
0	1	0	1	0	1	0	0	0	1	0	1	0	1	1	1	1	40
1	1	0	0	0	1	0	0	1	0	0	0	1	0	1	0	1	58
2	1	1	0	0	1	1	0	0	1	0	1	0	1	1	0	1	41
3	1	0	0	1	1	1	1	0	1	0	1	0	0	0	0	1	45
4	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	60

तालिका-4 : "आयु" कॉलम के साथ संधारित संख्यात्मक डेटा

[Table-4: Transformed numerical data set with 'Age' column]



चित्र-2 : प्रारंभिक मधुमेह भविष्यवाणी डेटासेट के लिए सहसंबंध मैट्रिक्स हीट-मैप

[Figure-2 : Correlation matrix heat-map for the early diabetes prediction dataset]

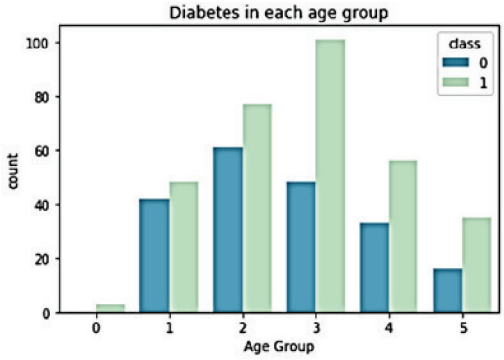
चित्र-2 से यह देखने योग्य है कि उम्र (0.11 सहसंबंध गुणांक) और मधुमेह के बीच कम सहसंबंध था। पॉल्यूरिया और पॉलीडिसेप्सिया, क्रमशः 0.67 और 0.65 के सहसंबंधों के साथ, मधुमेह के साथ उच्चतम संबंध थे। क्रमशः -0.01 और 0.05 के सहसंबंध के साथ, देरी से उपचार और खुजली का मधुमेह के साथ सबसे कम संबंध था। अधिकांश विशेषताओं में कम मधुमेह एसोसिएशन था। इसके अलावा, हमने उम्र के गुणों को श्रेणीबद्ध चर (1.15-25, 2.43, 3.36-45, 4.46-55, 5.56-65, 6 above 65) में बदल दिया और जांच की कि मधुमेह आयु समूहों से संबंधित है या नहीं जैसा कि तालिका-5 में दिखाया गया है।

	Gender	Polyuria	Polydipsia	sudden weight loss	weakness	Polyphagia	Genital thrush	Age Group	visual blurring	Itching	Irritability	delayed healing	partial paresis	muscle stiffness	Alopecia	Obesity	class	Age
0	1	0	1	0	1	0	0	3	0	1	0	1	0	1	1	1	1	40
1	1	0	0	0	1	0	0	5	1	0	0	0	1	0	1	0	1	58
2	1	1	0	0	1	1	0	3	0	1	0	1	0	1	1	0	1	41
3	1	0	0	1	1	1	1	3	0	1	0	1	0	0	0	0	1	45
4	1	1	1	1	1	1	0	5	1	1	1	1	1	1	1	1	1	60

तालिका-5 : आयु विशेषता श्रेणीगत चर में बदल गई

[Table-5 : The age attribute changed to categorical variables]

सबसे पहले, हम आयु समूह और मधुमेह के बीच एक संबंध है या नहीं यह जांचने के लिए आयु समूह के वितरण को देखने के लिए बार ग्राफ की साजिश करते हैं। आयु समूह की विशेषता के भीतर, हम तब मधुमेह के वितरण की साजिश करते हैं जैसा कि चित्र -3 में दिखाया गया है।

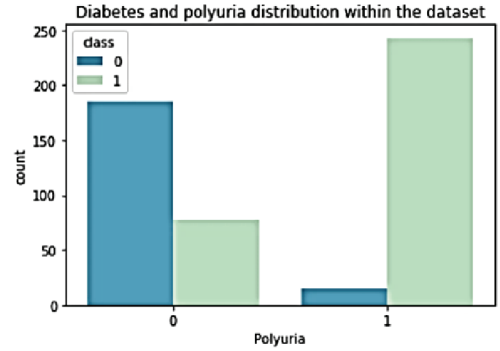


चित्रा-3 : आयु वर्ग और मधुमेह वर्ग वितरण
[Figure-3 : age group and diabetic class distribution]

Group आयु समूह 'और' वर्ग वितरण 'के प्रतिशत से, हम यह निष्कर्ष निकाल सकते हैं कि आयु समूह 4 (यानी, 46-55 वर्ष) में, रोगी में मधुमेह के लक्षण अधिक थे। जैसे, आयु समूह और मधुमेह के बीच संबंध सांख्यिकीय रूप से महत्वपूर्ण नहीं है क्योंकि मधुमेह सभी आयु समूहों में समान रूप से वितरित किया जाता है।

इसके बाद, आयु और मधुमेह के लिए ची-स्क्वायर परीक्षण किया गया, जिसके परिणामस्वरूप 0.076 पी-मान हुआ। पी-मानझ 0.05, इसलिए हम अशक्त परिकल्पना को नहीं तोड़ते हैं कि मधुमेह और आयु वर्ग के बीच कोई संबंध नहीं है।

प्राप्त किया गया पी-मूल्य 0.05 से कम था जब पॉल्यूरिया और मधुमेह के बीच सहयोग की जांच के लिए ची-स्क्वायर परीक्षण किया गया था। तो, हम अशक्त परिकल्पना को खारिज करते हैं और कहते हैं कि पॉलीयूरिया और मधुमेह का गंभीर संबंध है जैसा कि चित्र-4 में दिखाया गया है।



चित्र-4 : जिन रोगियों में पॉल्यूरिया था, उनमें मधुमेह का वितरण
[Figure-4 : Distribution of diabetes in patients who had polyuria]

3.4. अलग-अलग क्लासिफायर का उपयोग करके मधुमेह की भविष्यवाणी कर

मैंने नायव बेयस क्लासिफायर (एनबी), लॉजिस्टिक रिग्रेशन क्लासिफायर (एलआर), जे 48 एल्गोरिथम, रैंडम फॉरेस्ट (आरएफ) और मल्टी-लेयर परसेप्ट्रॉन (एमण्एलण्पीण्) एल्गोरिदम का उपयोग करके विश्लेषण किया है जैसा कि टेबल -6 में दिखाया गया है। और एमएलपी को दस गुना क्रॉस-वैलिडेशन और प्रतिशत स्प्लिट मूल्यांकन तकनीकों को लागू करने के बाद इस डेटासेट पर सबसे अच्छी सटीकता पाई गई। जटिल लक्षणों के पूर्वानुमान के लिए चिकित्सा क्षेत्र के भीतर व्यापक उपयोग के लिए, मल्टी-लेयर पर्सेप्ट्रॉन (एमएलपी) वास्तव में वरीयता का तंत्रिका नेटवर्क था। मॉडल की प्रभावशीलता को प्रशिक्षित करने और मूल्यांकन करने के लिए, डेटा को प्रशिक्षण और परीक्षण सेट में विभाजित किया गया था। हमने 80% डेटासेट को ट्रेनिंग सेट और अन्य 20% टेस्ट डेटा में विभाजित किया है।

Performance Parameters	Class	Algorithms NB	LR	J48	RF	MLP
Precision	WeightedAverage	0.879	0.924	0.957	0.974	0.985
Recall	WeightedAverage	0.874	0.924	0.956	0.974	0.980
F–measure	WeightedAverage	0.875	0.924	0.956	0.974	0.980

तालिका-6 : 10-गुना क्रॉस-सत्यापन का उपयोग करके प्रदर्शन की तुलना
[Table-6: Comparison of Performance using 10-fold Cross-validation]

रैंडम फ़ॉरेस्ट रीग्रेसर 97.4% सटीकता प्राप्त कर सकता है लेकिन हमारे एमण्एलण्पी. इस समान डाटासेट [dataset] के साथ 98% सटीकता प्राप्त करते हैं।

छिपी हुई परत के आकार को 12 नोड्स की 3 परतों में सेट किया गया था और हमारे डिजाइन द्व ारा 1000 के संचयी पुनरावृत्ति का उपयोग किया गया था। परीक्षण डेटा की भविष्यवाणी डेटा के प्रशिक्षण के बाद की गई और तकनीक की सफलता का आखिरकार विश्लेषण किया गया।

Confusion Matrix:

[44 2]
[0 58]

Classification Report:

	precision	recall	f1–score	suppor
0	1.00	0.96	0.98	46
1	0.97	1.00	0.98	58
accuracy			0.98	104
macro avg	0.98	0.98	0.98	104
weighted avg	0.98	0.98	0.98	104

हमने देखा कि 104 में से दो (2) रोगियों को भ्रम के मैट्रिक्स से गलत तरीके से लेबल किया गया था, जिसके परिणामस्वरूप 98 प्रतिशत सटीकता और 98 प्रतिशत एफ-स्कोर थे, जो मजबूत भविष्यवाणी उपाय हैं।

3.5. अनिश्चितता से निपटने के लिए प्रस्तावित कार्यप्रणाली

वर्तमान कार्य में, पीआईएम, मधुमेह डेटासेट की अनिश्चितता की अंतर्निहित डिग्री तक पहुंचने और मधुमेह की संभावना को निर्धारित करने के लिए एएनएफआईएस आधारित वर्गीकरण मॉडल के डिजाइन के बाद एक अस्पष्ट (फर्जी) आधारित विशेषज्ञ प्रणाली विकसित की गई थी।

इसके बाद, प्राप्त परिणामों की तुलना मल्टी-लेयर परसेप्ट्रॉन (एमएलपी) आधारित वर्गीकरण से की गई। परिणामों से पता चला कि एएनएफआईएस मॉडल बेहतर वर्गीकरण सटीकता प्रदान करता है।

विशेषज्ञ प्रणाली PIMA भारतीय मधुमेह डेटाबेस (PIDD) का उपयोग करके विकसित की गई थी। मधुमेह मेलिटस विभिन्न आयु समूहों के पुरुष और महिला दोनों को प्रभावित कर सकता है। मधुमेह की घटना की संभावना की गणना करने के लिए प्रत्येक रोगी से इनपुट के रूप में ग्लूकोज स्तर, इंसुलिन स्तर, बॉडी मास इंडेक्स (बीएमआई), मधुमेह वंशावली समारोह (डीपीएफ) और उम्र जैसे पैरामीटर दर्ज किए गए थे। इन परीक्षा परिणामों को अस्पष्ट (फजी) डेटा में परिवर्तित कर दिया गया था, जिसमें आउटपुट मान शून्य और एकता के बीच भाषाई चिकित्सा धारणाओं पर विचार कर रहे थे। तत्पश्चात, अस्पष्ट (फजी) मूल्यों से परिणाम प्राप्त करने के लिए ज्ञान आधारित निर्णय लिया गया। डिफ्यूज़िफिकेशन के साथ, फ़ज़ी आउटपुट वेरिएबल के मनमाने पैमाने पर क्रिस्प वैल्यू के रूप में अंतिम मान प्राप्त किए गए जो मधुमेह निदान की संभावना का वर्णन करता है।

3.5.1. फ़ज़ी आधारित विशेषज्ञ प्रणाली

प्रस्तावित प्रणाली में पहला कदम इनपुट और आउटपुट विशेषताओं को परिभाषित करना था। नैदानिक संभावना के आधार पर आउटपुट को शब्दार्थ रूप से बहुत कम, निम्न, मध्यम, उच्च, बहुत उच्च के रूप में व्यवस्थित किया गया था। सदस्यता मूल्यों को इनपुट विशेषताओं को अस्पष्ट (फजी) करने के लिए परिभाषित किया गया था। और ट्रैप-ज़ॉइडल सदस्यता फ़ंक्शन का उपयोग फ़ज़िफ़िकेशन करने के लिए किया गया था।

फ़ज़ी ऑपरेटर 'AND' या 'R' को मिसाल में लागू किया गया था और उसके बाद अगले चरण में अनुमान लगाया गया था यानी मिसाल से परिणामी तक और फिर परिणामी को पूरे परिभाषित नियमों में जोड़ा गया था। ये नियम चिकित्सा क्षेत्र के विशेषज्ञ द्वारा दिए गए फीडबैक पर आधारित हैं। इस उद्देश्य के लिए फ़ज़ी लॉजिक MATLAB टूलबॉक्स का उपयोग किया गया था।

अगला कदम सिंगल क्रिस्प आउटपुट वैल्यू प्राप्त करने के लिए सट्रोइड विधि का उपयोग करके डिफ्यूजिफिकेशन था। इस प्रकार परिणाम ने हम निदान की संभावना दी।

3.5.2. अनुकूली न्यूरो-फजी अनुमान प्रणाली

ANFIS का निर्माण खंड ताकागी-सुजेनो फजी इंट्रस सिस्टम [10] है। इस वास्तुकला में पाँच इनपुट विशेषताओं और एक आउटपुट की पाँच परत शामिल हैं। फ़ज़ीफ़िकेशन और नियम परत क्रमशः पहली और दूसरी परत हैं। चौथी परत तीसरी परत से इनपुट लेती है जिसने मूल्यों को सामान्य किया। डिफ्यूज़ी-फ़िकेटेड मान अंतिम परत को पास किए जाते हैं जो अंतिम आउटपुट देता है।

परत-I और परत-IV के कारक शिक्षार्थी प्रकार के थे। प्रथम परत कारक सदस्यता कार्यों को निर्धारित करता है जो इस मामले में ट्रैप-ज़ॉइडल था। हाइब्रिड एल्गोरिथ्म को प्रशिक्षण और अनुकूलन के लिए लागू किया गया था जो आउटपुट के गुणांक को अद्यतन करने के लिए कम से कम वर्ग अनुमान पद्धति का उपयोग करता है, इस प्रकार आधार और परिणामी मापदंडों को अनुकूलित करता है और बैक-प्रोपेगेशन एल्गोरिथ्म के विपरीत तेजी से परिवर्तित होता है जो केवल मौलिक मापदंडों को अपडेट करता है।

3.5.3. प्रशिक्षण और परीक्षण

प्रारंभ में, 80% PIDD मान इनपुट और आउटपुट वैक्टर दोनों के साथ नेटवर्क को प्रशिक्षित करने के लिए लोड किए गए थे। हाइब्रिड ऑप्टिमाइजेशन एल्गोरिथम को चुनकर ट्रेपोजॉइडल सदस्यता फ़ंक्शन 3-3-3-3-3 को चुनकर फजी इंट्रस सिस्टम तैयार किया गया था।

ANFIS मॉडल प्रशिक्षण पूरा होने के बाद, च्क्क् के शेष 20% का उपयोग मॉडल का परीक्षण करने के लिए किया गया था। प्रशिक्षण और परीक्षण प्रक्रिया दोनों के दौरान रूट मीन स्क्वायर एरर (आरएमएसई) मानों की गणना करके मॉडल के प्रदर्शन को सत्यापित किया गया था। विभिन्न युग मूल्यों के लिए प्रशिक्षण और परीक्षण प्रक्रिया दोनों को दोहराया गया था।

3.6. परिणाम

3.6.1. मधुमेह की संभावना

फजी मॉडल के परिणामों को मधुमेह की संभावना या गंभीरता के आधार पर पांच अलग-अलग समूहों में वर्गीकृत किया गया था यानी बहुत कम, निम्न, मध्यम, उच्च, बहुत अधिक।

3.6.2. हाइब्रिड एल्गोरिथम के साथANFIS मॉडल का प्रदर्शन

इसके अलावा, MATLAB फ़ज़ी लॉजिक टूलबॉक्स में उत्पन्न ANFIS मॉडल को प्रशिक्षित और परीक्षण किया गया था। मॉडल की प्रदर्शन दक्षता को विभिन्न EPOCH के लिए निर्धारित RMSE मूल्यों के माध्यम से मान्य किया गया था। अलग-अलग EPOCH में (जैसे, 10, 20, 50, 100, 150) हम प्रशिक्षण और परीक्षण दोनों चरणों के लिए एक ही आरएमएसई (0.2276) प्राप्त हुआ।

एक बहु-परत परसेप्ट्रॉन (MLP) तंत्रिका नेटवर्क को समान पीआईएम, डेटासेट को वर्गीकृत करने के लिए डिज़ाइन किया गया था। सटीकता, संवेदनशीलता और विशिष्टता जैसे सांख्यिकीय मापदंडों को भ्रम मैट्रिक्स उत्पन्न करके और ANFIS मॉडल से प्राप्त की तुलना में मापा गया था। परिणाम नीचे तालिका-7 में दर्शाए गए हैं।

Classification model → Performance Metrics ↓	MLP	ANFIS [10]
Sensitivity (%)	98.7	93
Specificity (%)	72.5	74.17
Accuracy (%)	85.3	87.24

तालिका 7 : विभिन्न वर्गीकरण मॉडल का प्रदर्शन सत्यापन
[Table-7: Performance Validation of Different Classification Models]

प्राप्त परिणामों से पता चला है कि MLP वर्गीकरण मॉडल की तुलना ANFIS वर्गीकरण मॉडल में बेहतर सटीकता थी।

4. निष्कर्ष

खंड 3.4 के भ्रम मैट्रिक्स से। 104 नमूनों में से केवल 2 रोगियों को गलत तरीके से वर्गीकृत किया गया था, जिसके परिणामस्वरूप 98% सटीकता और 98% प्रतिशत एफ-स्कोर था। ये सफल भविष्यवाणी के मार्कर हैं। मशीन सीखने की तकनीक का उपयोग करते हुए, मधुमेह के शुरुआती पता लगाने के लिए दोनों विशिष्ट और कम

विशिष्ट मधुमेह लक्षणों का उपयोग किया जा सकता है। इसकी सटीकता के कारण, डस्च् मशीन लर्निंग मॉडल एक उत्कृष्ट सूट है। तो, प्रारंभिक अवस्था में मधुमेह का अनुमान लगाने के लिए MLP मशीन लर्निंग मॉडल का उपयोग किया जा सकता है।

अनिश्चितता के अध्ययन में डेटा में अनिश्चितता और मॉडल में अनिश्चितता शामिल है। मापन शोर, संचरण शोर, और लापता मूल्यों जैसे स्रोतों से डेटा अनिश्चितता उत्पन्न होती है। मॉडल अनिश्चितता में सर्वोत्तम वास्तु. कला और मापदंडों को न जानना शामिल है जो भविष्य के डेटा की भविष्यवाणी कर सकते हैं। अनिश्चितता की मात्रा का ठहराव विभिन्न तरीकों से प्राप्त परिणामों में विश्वास बढ़ाने में मदद करता है। इसलिए, डेटा और मॉडल दोनों में अनिश्चितता से निपटना शोधकर्ताओं के लिए मेडिकल डोमेन में सटीक निर्णय लेने के लिए एक महत्वपूर्ण विषय है। फ़ज़ी सिस्टम सबसे व्यापक रूप से उपयोग की जाने वाली तकनीक है। ये विधियां सरल हैं और विभिन्न प्रकार की अनिश्चितताओं को कुशलता से दूर भी कर सकती हैं।

फजी आधारित विशेषज्ञ प्रणाली का उपयोग मधुमेह मेलिटस के प्रारंभिक चरण में होने की संभावना का पता लगाने के लिए किया गया है जैसा कि खंड ३ण्5 में चर्चा की गई है। इस पद्धति में मानव नियंत्रण तर्क का अनुकरण करने वाले सटीक और अस्पष्ट चिकित्सा डेटा को स्वीकार करके निर्णय लेने की क्षमता है और इस प्रकार पारंपरिक तरीकों के विपरीत उच्च सटीकता सुनिश्चित करता है। ANFIS को हाइब्रिड एल्गोरिथम के साथ लागू करके फजी सिस्टम को और बेहतर बनाया गया। खंड 3.6 में चर्चा किए गए परिणाम, MLP जैसे अन्य वर्गीकरण एल्गोरिदम की तुलना में बेहतर सटीकता पर च्ड। डेटासेट को वर्गीकृत करने के लिए साबित हुए। प्रस्तावित दृष्टिकोण भी लागत प्रभावी है और जटिलता को कम करता है। भविष्य में सक्षम निदान प्रदान करने के लिए मॉडल की तुलना विभिन्न मशीन लर्निंग तकनीकों से की जानी है।

अंग्रेजी शब्द और इसके अनुरूप हिंदी समकक्ष शब्द :

Technical Terms (English)	**संबंधित हिंदी शब्द**
Artificial Intelligence	कृत्रिम होशियारी
Logistic Regression	संभार तन्त्र परावर्तन
Classification	वर्गीकरण
Machine Learning Algorithm	मशीन लर्निंग कलन विधि
Multi Layer Perceptron	मल्टी लेयर परसेप्ट्रॉन

संदर्भ (References)

Datasets:

- https:@@archive.ics.uci.edu@ml@datasets@Early+stage+diabetes+risk+prediction+dataset.
- https:@@www.kaggle.com@uciml@pima&indians–diabetes–database

[1] Shetty, Deeraj, Kishor Rit, Sohail Shaikh, and Nikita Patil. "Diabetes disease prediction using data mining." In 2017 international conference on innovations in information, embedded and communication systems (ICIIECS), pp. 1–5. IEEE, 2017.

[2] Singh, Asmita, and R. Lakshmiganthan. "Impact of dffierent data types on classifier performance of random forest, naive bayes, and k–nearest neighbors algorithms." Int JAdv Comput Sci Appl; 8:1–10, 2018.

[3] Singh, D.A.A. G., E. Jebamalar Leavline, and B. Shanawaz Baig. "Diabetes prediction using medical data." Journal of Computational Intelligence in Bioinformatics 10, no. 1 (2017) : 1–8.

[4] Azrar, Amina, Y.Ali, M.Awais, and Khurram Zaheer. "Data mining models comparison for diabetes prediction." Int JAdv Comput Sci Appl 9, no. 8 (2018) : 320–323.

[5] Alam, Talha Mahboob, Muhammad Atif Iqbal, YasirAli, Abdul Wahab, Safdar izaz, Talha Imtiaz Baig, Ayaz Husin et al. "A model for early prediction of diabetes." Informatics in Medicine Unlocked 16 (2019): 100204.

[6] Islam, MM Faniqul, Rahatara Ferdousi, Sadikur Rahman, and Humayra Yasmin Bushra. "Likelihood prediction of diabetes at early stage using data mining techniques." In Computer Vision and Machine Intelligence in Medical ImageAnalysis, pp. 113–125. Springer, Singapore, 2020.

[7] Cobos, Leopoldo. "Unreliable hemoglobin A1C (HBA1C) in a patient with new onset diabetes after transplant (nodat)." Endocrine Practice 24 (2018): 43–44.

[8] Dorcely, Brenda, Karin Katz, Ram Jagannathan, Stephanie S. Chiang, Babajide Oluwadare, Ira J. Goldberg, and Michael Bergman. "Novel biomarkers for prediabetes, diabetes, and associated complications." Diabetes, metabolic syndrome and obesity: targets and therapy 10 (2017) : 345.

[9] Merad–Boudia, Hamza Nadjib, Majda Dali–Shi, Youcef Kachekouche, and Nouria Dennouni–Medjati. "Hematologic disorders during essential hypertension." Diabetes – Metabolic Syndrome: Clinical Research – Reviews 13, no. 2 (2019): 1575–1579.

[10] Karaboga, Dervis, and Ebubekir Kaya. "Adaptive network based fuzZy inference system (ANFIS) training approaches: a comprehensive survey." Artificial Intelligence Review 52, no. 4 (2019): 2263–2293.

[11] Toğaçar, Mesut, Burhan Ergen, and Zafer Cömert. "COVID–19 detection using deep learning models to exploit Social Mimic Optimization and structured chest X–ray images using fuzZy color and stacking approaches." Computers in biology and medicine 121 (2020): 103805.

[12] Rundo, Leonardo, Lucian Beer, Stephan Ursprung, Paula Martin–Gonzalez, Florian Markowetz, James D. Brenton, Mireia Crispin–Ortuzar, Evis Sala, and Ramona Woitek." Tissue–specific and interpretable sub–segmentation of whole tumour burden on CT images by unsupervised fuzZy clustering." Computers in biology and medicine 120 (2020): 103751.

[13] Singh, Prem Kumar. "Three–way fuzZy concept lattice representation using neutrosophic set." International Journal of Machine Learning and Cybernetics 8, no. 1 (2017): 69–79.

[14] Singh, Prem Kumar. "Medical diagnoses using three–way fuzZy concept lattice and their Euclidean distance." Computational and Applied Mathematics 37, no. 3 (2018): 3283–3306.

विज्ञान की तीन विधियाँ हैं - सिद्धान्त , प्रयोग और सिमुलेशन। विज्ञान की बहुत सारी परिकल्पनाएँ गलत हैं ; यह पूरी तरह ठीक है । ये (गलत परिकल्पनाएँ) ही सत्य-प्राप्ति के झरोखे हैं ।

हम किसी भी चीज को पूर्णत: ठीक तरीके से परिभाषित नहीं कर सकते । अगर ऐसा करने की कोशिश करें तो हम भी उसी वैचारिक पक्षाघात के शिकार हो जायेंगे जिसके शिकार दार्शनिक होते हैं।

— रिचर्ड फिलिप्स फ़ेनिमैन (भौतिकी में नोबेल पुरस्कार विजेयिता) (1965)