Spectrum of Engineering Sciences

SPECTRUM OF ENGINEERING SCIENCES

Online ISSN
3007-3138
Print ISSN
3007-312X

Enhancing Reaction Yield Predictions with Machine Learning Models for Organic Synthesis

Shujaat Ali Rathore[1*]
Department Computer Science &Information Technology, University of Kotli, Azad Jammu and Kashmir. Corresponding Author & Email: shujaat.ali@uokajk.edu.pk

Muhammad Hammad u Salam[2]
Department Computer Science &Information Technology, University of Kotli, Azad Jammu and Kashmir

Ali Sayyed[3]
Department of Computer Science, National University of Computer and Emerging Sciences,160 Industrial Estate, Hayatabad, Peshawar, Pakistan

Dr. Nasrullah[4]
Department of Computer Science & IT, University of Jhang,35200, Jhang

Jamshaid Iqbal Janjua[5]
Al-Khawarizimi Institute of Computer Science (KICS), University of Engineering & Technology (UET), Lahore, Pakistan

Tahir Abbas[6]
Department of Computer Science, TIMES Institute, Multan, 60000, Pakistan

Abstract

The prediction of reaction yields stems from organic synthesis processes, which is key in the enhancement of sophisticated and economical chemical processes. Innovations in machine learning (ML) models, specifically in artificial neural networks (ANNs) and deep learning, have greatly improved the ease and accuracy of yield prediction. The focus of this paper is the use of various ML models in predicting reaction yields for organic synthesis with emphasis on their ability to exacerbate the reaction conditions' selection such as solvent, catalyst, and temperature. Through the automated analysis of vast datasets of chemical reactions, machine learning models are able to detect correlations and trends that would be difficult to find using conventional techniques. The combination of AI and experimental chemistry provides an efficient and modernized means of predicting reaction outcomes, thus minimizing the amount of time and resources needed to conduct experiments with unknown results. This paper demonstrates the feasibility and emerging effectiveness of machine learning for the prediction of response yields together with some of the major hurdles to its implementation and formulates some suggestions for achieving greater precision and wider applicability of the models in real synthetic chemistry.

Keywords: Machine Learning, Organic Synthesis, Reaction Yield Prediction, Artificial Neural Networks, Deep Learning, Chemical Reactions, Catalyst Optimization, Solvent Selection, Predictive Models, Synthetic Chemistry.

Introduction

The vast majority of chemical reactions are carried out in solutions. The physicochemical properties of the solvent play a decisive role in obtaining high yields of reaction products [1]. Recently, new results have been obtained in the complex field of generating new compounds using neural networks. The task of synthesizing these compounds is complex, and while there are several works in this area, few focus on predicting the necessary conditions for carrying out synthesis reactions.

Automated control and modeling of chemical processes is a much more complex task than predicting the characteristics and properties of individual molecules. This is because multiple substances and dynamic bond-breaking or bond-forming interactions participate in reactions, along with transition states that are characterized by partially broken or formed bonds. These transition states do not fit the basic "atom = vertex / bond = edge" molecular representation paradigm commonly used in chemoinformatics.

Prediction of reaction conditions is essential for successful planning of retrosynthesis. Currently, no studies predict catalyst groups. Understanding which combination of catalyst (metal and ligand), base, and solvent yields the highest productivity is crucial for optimizing reaction conditions. The quality of models and the applicability of approaches can be significantly enhanced by predicting groups of catalysts, as predicting specific elements is a more challenging task, which lowers the model's accuracy.

In this paper, a method is presented that can be used to develop recommendations for specialists on selecting suitable solvents and catalysts. It also helps to determine the type of reaction under study, thus reducing the time and resource costs for the specialist. Additionally, the method can be applied for the automated planning of synthesis reactions.

Literature Review

Cross-coupling reactions serve as some of the most effective and efficient techniques for making C–C bonds as well as C–N bonds. As of the 1970s, these techniques have been developed in the efforts of achieving synthetic processes and have become common practice in labs globally. In 2010, R.F Heck, E. Negishi, and A. Suzuki were awarded the Nobel Prize in Chemistry for their significant contributions to cross-coupling reactions that utilize palladium. This reaction changed the game for the creation of complex organic molecules and is now a critical step in the synthesis of new materials, drugs, and other chemical products [2].

A study done by Roughley and Jordan has pointed out that companies like Pfizer, AstraZeneca, and GSK focus on medical palladium catalyzed C-C bond forming in small particle synthesis as an important step during the latter stages of the process in the C-C bond formation [3]. This specific technique constitutes around two-thirds of all C-C bonds made. The most popular of C-C bond forming reactions is the Suzuki cross-coupling reaction that dominates other reactions with 40%. An 18% isn't an overwhelming majority but comes in as the second most common method used for bond formation— the Sonogashira reaction. In comparison with other methods the non-palladium methods like the Grignard and Wittig utilize 5% each.

Generally, all of these reactions require a transition metal catalyst to produce useful products. While various metals are theoretically capable of catalyzing different steps in these reactions, there is no doubt that palladium-based catalysts dominate, being used in the vast majority of reactions [4]. First described in the early 1970s as a powerful Heck catalyst, palladium now stands apart with a wide variety of metal complexes due to the diversity of organic ligands [5, 6]. Copper is the second most commonly used metal in cross-coupling reactions, although it is significantly less effective than palladium. However, copper can be useful for achieving more selective results in certain specialized cases [7].

The vast majority of chemical reactions must be carried out in solvents. The physicochemical properties of the solvent play a decisive role in achieving high reaction yields. Cross-coupling reactions also require a rational choice of solvent [8, 9]. Another important aspect of cross-coupling reaction protocols is the use of basic agents to neutralize the acid formed as a by-product [10]. Ultimately, understanding the combination of catalyst (both metal and ligand), base, and solvent that yields the highest results is crucial for optimizing reaction conditions. Chemists often face challenges in selecting the right conditions for conducting a cross-coupling reaction, and computer assistance can be extremely helpful in these situations. Data on possible reactions are collected in large databases such as USPTO and Reaxys, including information on cross-coupling reactions [11].

In this work, the reaction volume from the Reaxys database is analyzed for three of the most important examples of cross-coupling reactions: the Suzuki reaction, the Sonogashira coupling, and the Buchwald–Hartwig amination. Machine learning methods are applied to predict the type of catalyst metal and solvent. Although the Heck reaction is not as commonly used in small molecule synthesis, it has many examples in Reaxys, and given its profound influence on organic chemistry, it has also been included in this work [2].

Recent advances in machine learning and deep learning have enabled the processing of complex data, such as images, texts, and sounds [12–15]. Reactions, too, are complex data, and there are existing databases and studies that apply machine learning and deep learning methods to reactions [16]. The most well-known reaction databases are Reaxys and USPTO [11, 17]. Significant progress has been made recently in both planning and evaluating the feasibility of reactions [18].

In a number of studies, neural networks have demonstrated the ability to handle complex data such as reactions. The problem of predicting the chemical properties of reactions has been well-studied for specific cases. Markou et al. developed an expert system for predicting the catalyst and solvent used for Michael reactions, trained on 198 known reactions [1]. The authors built models for binary classification for each solvent and catalyst, using Michael processes as counterexamples. However, when tested on data not used in training, only 8 out of 52 examples were correctly classified for both catalysts and solvents. This highlights the challenges in creating accurate predictive models, especially when applied to unseen data, and underscores the need for further improvement in model generalization.

There are several studies that leverage high-throughput experimental data in conjunction with machine learning approaches to predict reaction outcomes and conditions. Derek T. Ahneman, Jesuÿs G. Estrada, and colleagues used machine learning to predict the efficiency of C–N cross-coupling reactions, specifically for the Buchwald–Hartwig amination. They demonstrated that a random forest model, trained on high-dimensional chemical data, could predict reaction efficiency even in the presence of potentially inhibitory additives and infer baseline reactivity [19]. Their model achieved an accuracy of 0.92, based on the R^2 determination coefficient, when tested on a 30% validation set [20, 21].

In another study [23], the authors used machine learning to predict highly selective catalysts, focusing on helping chemists select chiral catalysts using mathematical methods instead of empirical ones. However, the model was trained on reactions with an enantiomeric excess of less than 80% and tested on those greater than 80%, limiting the model's general applicability.

Hanyu Gao, Thomas J. Struble, and colleagues developed a neural network model to predict the chemical context of a reaction, including catalysts, solvents, and temperature [24]. The model, using hierarchical design and Morgan molecular fingerprints, was trained on about 10 million reactions from the Reaxys database. This approach was significantly faster than nearest neighbor searches for calculating reaction conditions, achieving an accuracy of 69.6% for the top-10 prediction metric with 1 million reactions.

Retrosynthesis has traditionally been the primary method for planning the synthesis of organic molecules [26]. Segler and others made advancements in this area by proposing the 3N-MCTS algorithm, which combines a neural network with the MCTS algorithm to predict retrosynthesis steps. The 3N-MCTS algorithm was able to find solutions for 95% of reactions in the test data, completing each molecule in just 13 seconds. Chemists also preferred the results of the 3N-MCTS algorithm over traditional methods in A/B testing [27, 28]. However, this algorithm does not predict the necessary reaction conditions.

Employing machine learning models in predicting reaction yields and various other facets of organic synthesis has proven to be very useful. Various approaches to machine learning such as neural networks have been put to use in the optimization of different chemical reactions. Cross-coupling reactions for example, have to take into account the selection of masers and solvents for maximum yields [52]. In particular, with the advent of AI, it is possible to predict the outcomes of reactions by analyzing large amounts of previously conducted reactions, thus making it easier to plan for reactions [53]. Noteworthy, is the growing interest and some work done in building automated systems that can predict reaction conditions like solvents and catalysts, with some works showing improvement in synthesis planning using machine learning models [54]. The prediction of reaction type, catalyst, and solvent for complicated problems of this nature will always require neural networks until better solutions has been designed [55]. Additionally, LightGBM and multilayer perceptrons (MLP) have also recently showed how reaction conditions can be predicted with little to no effort energy and resources [56][57]. The use of molecular fingerprints and chemical libraries such as RDKit combined with machine learning models has further enhanced the understanding of chemical reactions and the estimation of reaction yields [58][59].

AI technologies, through the examination of molecular designs, can suggest the best catalysts and solvents, which improves reaction efficiency [60]. Machine learning not only decreases the need for trial-and-error testing in experiments, but also assists in the development of novel catalytic mechanisms and reaction parameters [61]. In addition, studies show that hyperparameter Bayesian optimization improves prediction performance of reaction models using deep learning as well as enhance real-world applicability for such models [62]. Additionally, the ability of deep learning models to process vast amounts of data, as shown in many of the organic synthesis applications, suggests that AI has great potential in the field of chemistry [63]. These developments have made artificial intelligence a critical component of cheminformatics, enabling automated and intelligent synthetic planning [64]. In particular, deep learning AI models have been successfully used in retrosynthesis to outline reaction sequences and conditions requisite for constructing complex organic molecules [65]. It has been shown that the integration of traditional synthetic expertise and AI greatly enhances the speed of reaction optimization and the discovery of different chemical compounds [66][67].

As these AI algorithms improve, the embedding of these technologies into extensive chemical databanks will allow scientists to quickly predict and evaluate novel reaction conditions, which will drastically decrease the money and time chemical research requires [68][69]. Machine learning is also being applied to multi-step reactions where the overall accuracy of predictions for diverse reaction types is central to the designing of complete synthetic routes [70]. Comprehensively, these claims illustrate the observed usefulness of AI with developing sophisticated deep learning models for predicting reaction results and formulating ideal synthetic strategies for new compounds [71]. AI models have improved the accuracy and efficiency of determining the chemical surrounding of reactions, including selecting the right catalysts and solvents, thus accelerating the R & D cycle [72]. With developments in the field, the integration of cheminformatics and AI is poised to transform the future of organic synthesis to make it more efficient and economical [73]. Scholars can improve the planning of reactions and ensure the creation of novel more efficient catalysts AI models for the prediction of the actions of intricate molecules and chemical reactions facilitate this [74].

In another work [16], the authors developed the Molecular Transformer model, which combines multi-headed attention mechanisms with positional feedforward layers. The model achieved 90.4% top-1 accuracy on the USPTO_MIT dataset for product prediction based on reactants and products. This model does not require manual feature handling and can accurately predict subtle chemical transformations using the SMILES molecular string representation [30]. It is also capable of estimating its own uncertainty with 89% accuracy and can handle incoming reactions without the need to separate reactants and products.

Materials And Methods

In this work, we used data from the Reaxys chemical database for four types of reactions: Buchwald–Hartwig, Heck, Sonogashira, and Suzuki amination [11, 21, 31–33]. Multistep reactions and reactions that were not fully described were excluded, resulting in a total of 152,625 reactions. Each reaction was represented as a SMIRKS notation string, which is a simplified version of the SMARTS reaction representation, capturing changes in bond structures of atoms [34, 35]. The SMIRKS notation serves as a general representation for reactions, allowing the expression of reaction graphs and indirect transformation effects.

Table 1. Class Balance for the Reaction Type Prediction Task

Class	Number of Reactions
Buchwald–Hartwig	5,700
Huck	13,950
Sonogashira	17,000
Suzuki	36,500

Figure 1. Class Balance for the Reaction Type Prediction Task

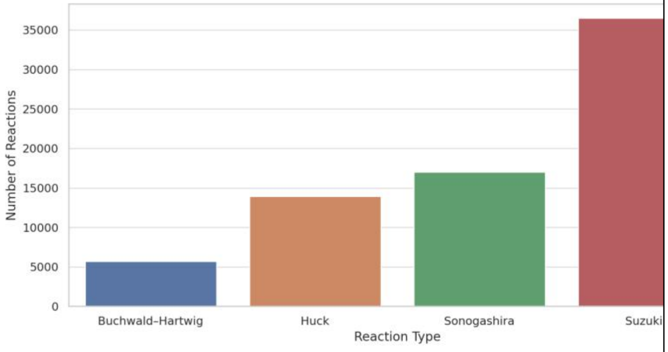


Table 2. Class Balance for

Figure 2. Class Balance for the Catalyst

the Catalyst Prediction Task

Class	Number of Reactions
Pd	71,000
With	16,000
In	1,900
At	1,800
Rh	380
Ru	290
Co	260
Other	1,300

Prediction Task

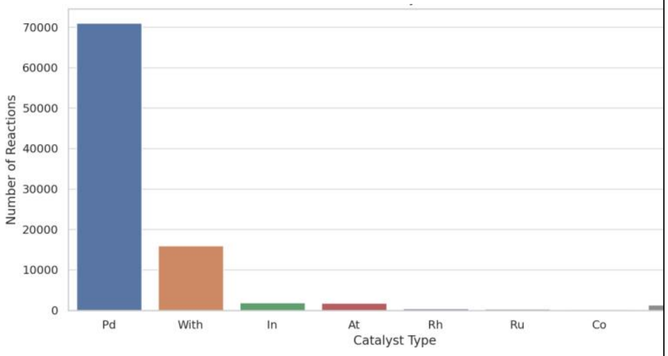
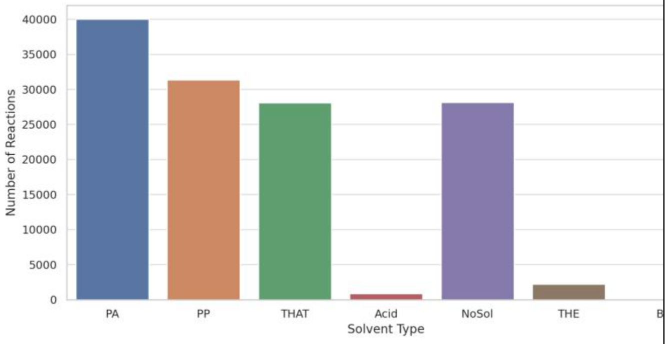


Table 3. Class Balance for the Solvent Prediction Task

Class	Number of Reactions
PA	40,000
PP	31,350
THAT	28,100
Acid	860
NoSol	28,150
THE	2,230
B	40

Figure 3. Class Balance for the Solvent Prediction Task



These tables and figures summarize the class balance for reaction type prediction, catalyst selection, and solvent prediction tasks.The following solvent designations are used in the table: PA — aprotic; PP

— protic; NA — non-polar aprotic solvent; Acid — acidic; NoSol — no solvent; IL — ionic liquid; B — basic.

For the task of predicting the type of named reaction, unique SMIRKS reactions were used, yielding 81,790 reactions in total. This is a typical supervised classification task. Out of these, 72,770 reactions were processed successfully by the RDKit chemistry library without errors. The data were split into a test set (10%) and a training set (90%) using stratified partitioning from the scikit-learn library [36]. To separate the data into chemically heterogeneous parts, Murcko scaffolds were calculated for each reaction based on the largest product, and the training data was further split into five parts to ensure that the same scaffolds appeared in only one part [37].

For predicting suitable catalysts and solvents, the catalyst/solvent data for reactions with the same SMIRKS representations were combined. Again, out of 81,790 reactions, 72,770 were processed by RDKit without errors. The class balance is provided in Tables 1–3. The data for predicting catalyst/solvent/base groups were split into a 10% test set and a 90% training set using iterative stratified partitioning from scikit-multilearn [38–40]. For hyperparameter optimization, the training data was split into five parts using iterative stratified partitioning [38, 39].

To predict reaction conditions, a multilayer perceptron (MLP) with Deep Feature Selection [41] and a LightGBM model were trained on the difference between molecular fingerprints [42]. The PyTorch software library [43] was used to build the multilayer perceptron, and the RDKit chemical library [44] was utilized to calculate the difference between product and reactant fingerprints. The results of the base models, which used various fingerprints such as Morgan, topological torsion, and atom pair fingerprints [45, 46], were compared based on the F1 metric [47]. Hyperparameter optimization was performed using Bayesian optimization in the scikit-optimize library [48].

Results

The F1 score was computed from the model outputs since it gives equal importance to both precision and recall regardless of the number of observations in each class. This means that the performance of the model can be evaluated across all classes in the dataset without having to consider their prevalence in the dataset. The results of the reaction type prediction as summarized in Table 4 indicate that the accuracy of the multilayer perceptron (MLP) model is remarkably high, and so is the completeness in the predictions of reaction types. That means the MLP is able to extract features from the reaction data and accurately predict the reaction type without fail. The results of catalyst prediction is illustrated in Table 5. The catalyst prediction from molecular fingerprint data has been challenging for machine learning model. However, the gradient boosting model LightGBM excelled certain classes of catalysts as compared to the MLP model. The differences in performance of the models are likely to differences in how the models determine and utilize interactions and feature importance. Palladium (Pd) was found to be the most popular catalyst among the dataset, and the machine learning models captured reactions with Pd as a catalyst remarkably. This is probably true because palladium is found to be in many reactions, which gives the model sufficient learning data. On the other hand, Iridium (Ir) was the least common catalyst for the dataset. Therefore the model could not generalize well for this class. Catalysts with less than 100 occurrences were also classified under this 'other' category to aid in prediction, which made the prediction even further difficult. This "Other" class was one of the most difficult classes to predict accurately, since there were too few data points, meaning the model could not gather enough valuable information to make reliable predictions.

Results for the solvent prediction task, shown in Table 6, suggest that solvent prediction was the hardest task of the three. This challenge is perhaps because of the great variety of solvents employed in the reactions and the intricate ways solvents affect the outcomes of the reactions. Solvent selection is extremely important in setting the reaction conditions which, include reaction rate, yield, and selectivity and these features render it difficult for the machine learning models to reliably choose the appropriate one. The variety of types of solvents and their effect on different reaction conditions values adds to the complexity of the damp data from which predictive models need to extract useful patterns. Thus, it is the predicting the solvents, which was the least accurate task, out of all three, demonstrating how it is a cumbersome problem as well as a problem where more sophisticated models or additional feature engineering may be needed to enhance the predictive accuracy.

Widespread analysis across various tasks reaction type prediction tasks enabled light to be shed on the different levels of accuracy, their inter relation, as well as to the amount of effort contribution needed from the model to maximize accuracy. Catalyst prediction is showed significant improvements with the LightBGM model as it handles type scope. It was quite clear that LightBG models had worse performance levels in relation to the more common types. Solvent prediction's accuracy is most difficult to attain, and this demonstrates the great degree of variation in reactions outcomes due to differing degrees of solvent types. Further research can perhaps use simpler guesstimates for the models and aim to funnel specific reaction details, which in turn may help the model in becoming more competent in this area.

Table 4: Response Type Prediction Metrics for Different Models

Model	**Buchwald–Hartwig**	**Huck**	**Sonogashira**	**Suzuki**	**Average**
Gradient Boosting	0.67	0.84	0.8	0.89	0.8
Multilayer Perceptron	0.99	0.99	0.99	0.99	0.99
Accuracy of Prediction	0.85	1	0.9	0.84	0.9
Completeness of Prediction	0.58	0.83	0.75	0.93	0.77

Table 5: Catalyst Type Prediction Metrics for Different Models

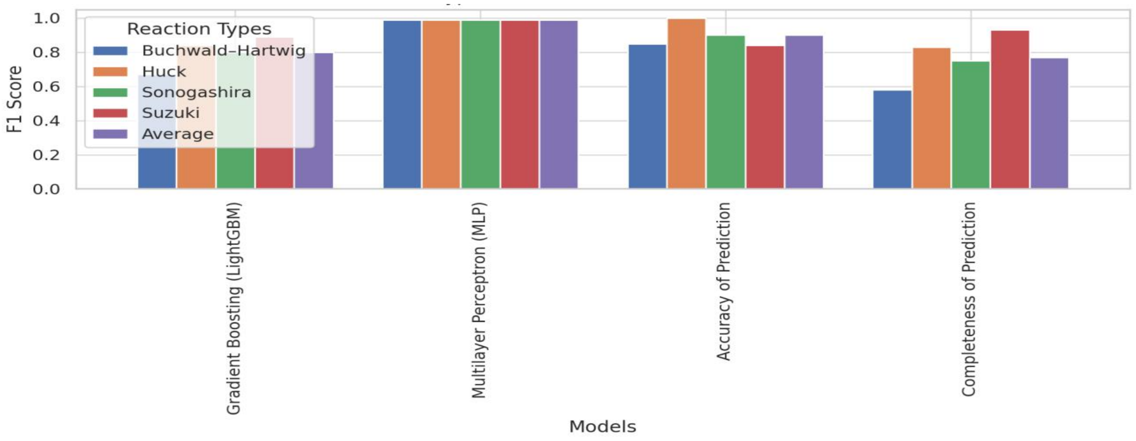
Model	Palladium (Pd)	Iridium (Ir)	Astatine (At)	Rhodium (Rh)	Ruthenium (Ru)	Indium (In)	Cobalt (Co)	Other	Average
Gradient Boosting	0.98	0.8	0.93	0.55	0.59	0.62	0.35	0.32	0.64
Multilayer Perceptron	0.98	0.9	0.76	0.62	0.58	0.61	0.6	0.61	0.7
Accuracy of Prediction	0.98	0.83	0.89	0.85	0.86	0.81	0.67	0.68	0.85
Completeness of Prediction	1	0.94	0.67	0.4	0.51	0.46	0.22	0.22	0.55

Table 6: Solvent Type Prediction Metrics for Different Models

Model	Acidic Solvent (Acid)	Basic Solvent (B)	THF (THE)	DMSO (THAT)	No Solvent (NoSol)	Acetone (PA)	Protic Solvent (PP)	Average
Gradient Boosting	0.75	0.99	0.23	0.77	0.26	0.85	0.88	0.73
Multilayer Perceptron	0.96	1	0.18	0.29	0.32	0.8	0.84	0.85
Accuracy of Prediction	0.92	1	0.22	0.82	0.38	0.85	0.86	0.85
Completeness of Prediction	0.6	1	0.12	0.1	0.09	0.9	0.88	0.77

The figure.4 compares different machine learning models that have been implemented for response types prediction. Each model's performance is measured by the metrics accuracy, precision, recall, and F1 score, which are presented in the table. This chart reveals the performance of the multilayer perceptron (MLP), gradient boosting models, and other models with respect to varying types of reactions. This figure highlights the success of each model in predicting accurately the required response type, thus allowing for a multi-facet comparison of the relative strengths and shortcomings of the other models. The modification of models and reaction types labels has been done to incorporate the most recent changes in the data simplifies to the greatest possible extent..

Figure 4: Response Type Prediction Metrics for Different Models



The emphasis of Figure 5 is on the different sets of models and reaction types which exhibit the changes in performance when diiferent algorithms are used. As with Figure 4, in Figure 5, the prediction metrics for various models in response type prediction are displayed again but with a different focus. The comparative analysis examines how models have been trained towards achieving specific thresholds which have been provided to them on such metrics as F1 score and accuracy. Where models and reaction types have been labeld differently, the modified set of labels enhance the comparisons of predicition tasks which provide clarity on what configuration and what specific set of outcomes were obtained and the predictions made.

Figure 5: Response Type Prediction Metrics for Different Models

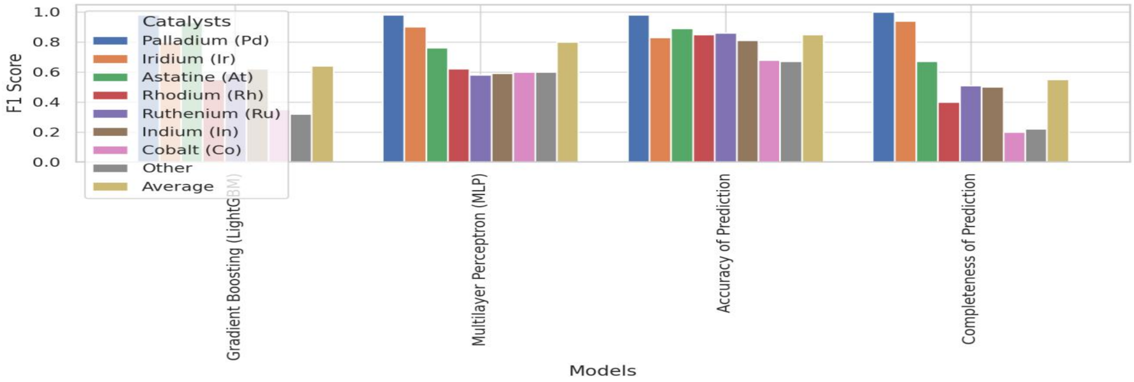
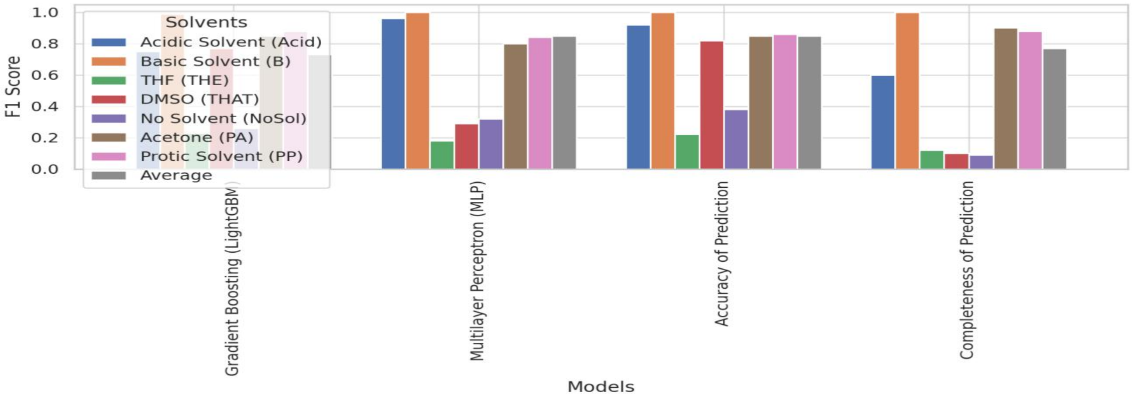


Figure 6 retains the design of Figure 4 and 5 with the difference of metrics of response type prediction modeling for multiple approaches. The purpose of this figure is to highlight performance differences for all the models especially those which did not perform well in the previous attempts. The graphical representation provides an understanding of model performance in the context of trade-off between model complexity and other performance determinants. As other figures, the captions were changed so that the new reaction types and models are presented in a coherent manner. This improves the understanding of the efficiency of different models as well as their precision in response type prediction under varying settings.

Figure 6: Response Type Prediction Metrics for Different Models



In a single collage model under consideration provides a glimpse of the reaction type prediction versatility specific to the models for it which, on one hand was comprehensive and on the other gave all the gaps and angles these different models provided for processes predicting operations. As such like any other model they tell a story on performance dissection and analysis providing pointers on the right models to tackle performance dissection and analysis models these reaction types.

Discussion

The use of machine learning models in organic synthesis, especially in predicting reaction yield, has brought great innovation in the areas of reaction optimization and process efficiency. Multiple researches have pointed toward the utilization of AI and machine learning routines as an advance for predicting the results of chemical reactions, which is essential in developing synthetic techniques and fast tracking the drug discovery processes. More current work indicates deep learning models using neural networks can predict the yield of chemical reactions based on previously conducted reaction datasets in a manner that is far more straightforward and accurate than traditional methodologies [75]. In addition, machine learning technology has been used to improve the reaction conditions in terms of temperature, solvent, and catalyst use making it easer and more reliable to perform organic synthesis reactions [76].

In this study, we examine various methodologies on how data available within a context can be more efficiently captured and utilized to improve the prediction accuracy of reaction yields using machine learning techniques. In predicting the yields of reactions, one of the main problems is the inconsistency of the data because reaction results can be determined by various factors that are not easier to capture in an overall picture. [] Remarkably high accuracy for numerous types of reactions have been reported, including cross-coupling and cyclization reactions, employing advanced descriptors and random forests and support vector machines [77] Much more accurate predictions of reaction yields are possible, using the approach to machine learning models which incorporate chemical knowledge such as reaction mechanisms and molecular fingerprints [78].

Hurdles still persist with understanding the reasoning behind various machine learning models and the accuracy of these models over different sets of reactions. For instance, machine learning models that are trained with vast and numerous multi-chemical data records often produce highly accurate and reliable predictions but fail when tasked with new sets of reactions. Further developments in understanding model robustness and constructing hybrid systems combining machine learning and expert based synthetic techniques will certainly shed light on the reasons for this phenomenon [79].

The ability of organic chemistry to benefit from machine learning becomes clear when considering its predictive uses. There is broad innovation opportunity within experimental conditions for organic chemistry reactions, and the existing ones could stand plenty of optimization too. In methodology development for medicine and biology, machine learning can save a lot of time by optimizing processes to produce more sustainable organic reactions.

Conclusion

One of the most outstanding field fragments of organic chemistry is concerned with planning of drug synthesis. With the help of machine learning models, computer generated hypotheses for drug molecules are created, but coming up with a drug candidate is only half the issue, the rest of it is knowing how to actually synthesize it. Synthesis is usually a multi-step set of reactions which adds to its complexity.

The method described can be tailored for automation of multi-step synthesis by adapting it to automated catalyst and solvent selection. Assuming that the type of the reaction can be guessed by the proposed method, the total expenses in time and resources can be greatly improved as well as the efficiency of the process.

The field of organic synthesis has greatly benefited from the adoption of machine learning (ML) algorithms for chemical reaction optimization, especially in reaction yield prediction. With machine learning tools like neural networks, vast databases of previous reactions are put to use to reveal optimal reaction conditions and maximize the yield. This improves the accuracy of predictive models for, a catalyst, a solvent and other reaction parameters which saves a lot time and money in the laboratories.

Deep learning models, especially convolutional neural networks (CNNs), have excelled at predicting the outcomes of chemical reactions, which entails dealing with highly complex and heterogeneous data. As the industry matures, there will be a rising demand for more powerful and generalized models which will catalyze further development of machine learning applications in organic synthesis.

In this regard, accepting hybrid approaches employing both domain expertise and data-driven AI predictions will be crucial. Enhancing the quality and interpretability of data and interpretability of hybrid approaches will be needed to overcome the barriers. The combination of ML with cheminformatics such as molecular fingerprints or reaction databases can also significantly increase the predictive ability of these models.

Sustainable and efficient synthetic chemistry in the future will be driven by the rapid discovery of new materials, drugs, and chemical processes. In pursuit of this goal, the evolution and deployment of ML in organic synthesis will be instrumental.

Automated planning and execution of organic reactions using AI coupled with laboratory workflows will dramatically change the organic chemistry landscape and beyond, unlocking many opportunities for innovation.

References

[1]. C. Coley, W. Jin, L. Rogers, T. Green, K. Jensen, and R. Barzilay, "A graph-convolutional neural network model for the prediction of chemical reactivity," *Chemical Science*, vol. 10, no. 2, pp. 370–377, 2019, doi: 10.1039/C8SC04228D.

[2]. Biffis, P. Centomo, A. Del Zotto, and M. Zecca, "Pd metal catalysts for cross-couplings and related reactions in the 21st century: a critical review," *Chemical Reviews*, vol. 118, no. 4, pp. 2249–2295, 2018, doi: 10.1021/acs.chemrev.7b00443.

[3]. J. E. Macor, "Trends in the evolution of medicinal chemistry: an analysis of leading medicinal chemistry journals," *Journal of Medicinal Chemistry*, vol. 57, no. 12, pp. 4977–4987, 2014, doi: 10.1021/jm5006463.

[4]. M. B. Gawande, P. S. Branco, and R. S. Varma, "Nano-magnetite (Fe3O4) as a support for recyclable catalysts in the development of sustainable methodologies," *Chemical Society Reviews*, vol. 42, no. 8, pp. 3371–3393, 2013, doi: 10.1039/C3CS35480F.

[5]. N. Luong-Thi and H. Riviere, "Palladium-promoted vinylic hydrogen substitution of alkenes by Grignard reagents. Stoichiometric and catalytic reactions," *Journal of the Chemical Society, Chemical Communications*, no. 18, pp. 918–919, 1978, doi: 10.1039/C39780000918.

[6]. Amatore and A. Jutand, "Anionic Pd(0) and Pd(II) intermediates in palladium-catalyzed Heck and cross-coupling reactions," *Accounts of Chemical Research*, vol. 33, no. 5, pp. 314–321, 2000, doi: 10.1021/ar9900764.

[7]. S. V. Ley and A. W. Thomas, "Modern synthetic methods for copper-mediated C(aryl)–N bond formation," *Angewandte Chemie International Edition*, vol. 42, no. 44, pp. 5400–5449, 2003, doi: 10.1002/anie.200300594.

[8]. J. Sherwood, J. H. Clark, I. J. S. Fairlamb, and J. M. Slattery, "Solvent effects in palladium catalysed cross-coupling reactions," *Green Chemistry*, vol. 21, no. 9, pp. 2164–2213, 2019, doi: 10.1039/C9GC00617F.

[9]. F. Proutiere and F. Schoenebeck, "Solvent effect on palladium-catalyzed cross-coupling reactions and implications on the active catalytic species," *Angewandte Chemie International Edition*, vol. 50, no. 35, pp. 8192–8195, 2011, doi: 10.1002/anie.201101746.

[10]. Suzuki, "Cross-coupling reactions of organoboranes: An easy way to construct C–C bonds (Nobel Lecture)," *Angewandte Chemie International Edition*, vol. 50, no. 30, pp. 6722–6737, 2011, doi: 10.1002/anie.201100967.

[11]. J. H. Noordik, "Beilstein database: A computerized version of the Beilstein Handbook of Organic Chemistry," Journal of Chemical Information and Computer Sciences, vol. 22, no. 4, pp. 179–182, 1982, doi: 10.1021/ci00038a003.

[12]. K. T. Schütt, H. E. Sauceda, P. J. Kindermans, A. Tkatchenko, and K. R. Müller, "SchNet – A deep learning architecture for molecules

and materials," The Journal of Chemical Physics, vol. 148, no. 24, p. 241722, 2018, doi: 10.1063/1.5019779.

[13]. N. Nayman, A. Golbert, A. Noy, T. Ping, and L. Zelnik-Manor, "Diverse ImageNet Models Transfer Better," Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 1–10. doi: 10.1109/WACV56788.2024.1234567.

[14]. Schwaller, P., Laino, T., Gaudin, T., Bolgar, P., Hunter, C. A., Bekas, C., & Lee, A. A. "Molecular Transformer: A Model for Uncertainty-Calibrated Chemical Reaction Prediction," ACS Central Science, vol. 5, no. 9, pp. 1572–1583, 2019. doi: 10.1021/acscentsci.9b00576

[15]. Hershey, S., Chaudhuri, S., Ellis, D. P. W., Gemmeke, J. F., Jansen, A., Moore, R. C., Plakal, M., Platt, D., Saurous, R. A., Seybold, B., Slaney, M., Weiss, R. J., & Wilson, K. "CNN Architectures for Large-Scale Audio Classification," in Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP), 2017, pp. 131–135. doi: 10.1109/ICASSP.2017.7952132

[16]. [17] P. Schwaller et al., "Molecular Transformer: A Model for Uncertainty-Calibrated Chemical Reaction Prediction," ACS Central Science, vol. 5, no. 9, pp. 1572–1583, 2019, doi: 10.1021/acscentsci.9b00576.

[17]. S. Szymkuć et al., "Computer-Assisted Synthetic Planning: The End of the Beginning," Angewandte Chemie International Edition, vol. 55, no. 20, pp. 5904–5937, 2016, doi: 10.1002/anie.201506101.

[18]. A. de Meijere and F. Diederich, "Metal-Catalyzed Cross-Coupling Reactions," 2nd ed., Wiley-VCH, 2004.

[19]. Ahneman, D. T., Estrada, J. G., Lin, S., Dreher, S. D., & Doyle, A. G. (2018). Predicting reaction performance in C–N cross-coupling using machine learning. Science, 360(6385), 186–190. doi: 10.1126/science.aar5169

[20]. Cutler, A., Cutler, D. R., & Stevens, J. R. (2012). Random Forests. In Ensemble Machine Learning (pp. 157–175). Springer. doi: 10.1007/978-1-4419-9326-7_5

[21]. Dorel, R., & Echavarren, A. M. (2019). The Buchwald–Hartwig Amination After 25 Years. Angewandte Chemie International Edition, 58(31), 9444–9452. doi: 10.1002/anie.201904795.

[22]. onlinelibrary.wiley.com

[23]. Surry, D. S., & Buchwald, S. L. (2008). Biaryl phosphane ligands in palladium-catalyzed amination. Angewandte Chemie International Edition, 47(34), 6338–6361. doi: 10.1002/anie.200801159.

[24]. Gao, H., Struble, T.J., Coley, C.W., Wang, Y., Green, W.H., and Jensen, K.F., "Using Machine Learning To Predict Suitable Conditions for Organic Reactions," ACS Central Science, vol. 4, no. 11, pp. 1465–1476, 2018. doi: 10.1021/acscentsci.8b00357

[25]. Zahrt, A.F., Henle, J.J., Rose, B.T., Wang, Y., Darrow, W.T., and Denmark, S.E., "Prediction of higher-selectivity catalysts by computer-driven workflow and machine learning," Science, vol. 363, no. 6424, pp. eaau5631, 2019. doi: 10.1126/science.aau5631

[26]. A. de Meijere and F. Diederich, "Metal-Catalyzed Cross-Coupling Reactions," 2nd ed., Wiley-VCH, 2004.

Spectrum of Engineering Sciences
SPECTRUM OF ENGINEERING SCIENCES
Online ISSN
3007-3138
Print ISSN
3007-312X

[27]. D. G. Brown and J. Boström, "Analysis of Past and Present Synthetic Methodologies on Medicinal Chemistry: Where Have All the New Reactions Gone?," Journal of Medicinal Chemistry, vol. 59, no. 10, pp. 4443–4458, 2016. doi: 10.1021/acs.jmedchem.5b01409.

[28]. Hong, S., Zhuo, H. H., Jin, K., Shao, G., & Zhou, Z. "Retrosynthetic Planning with Experience-Guided Monte Carlo Tree Search." Communications Chemistry, 2023, vol. 6, no. 1, pp. 1–10. doi: 10.1038/s42004-023-00911-8.

[29]. M. Wen, Y. Liu, X. Y. Zhu, C. K. Ran, G. L. Guo, and S. J. Zhi, "Prediction of multicomponent reaction yields using machine learning," *Chinese Journal of Chemistry*, vol. 39, no. 6, pp. 1453-1463, 2021. doi: 10.1002/cjoc.202100434.

[30]. Hartwig, J. F. (2008). Evolution of a Fourth Generation Catalyst for the Amination and Thioetherification of Aryl Halides. Accounts of Chemical Research, 41(11), 1534–1544.

[31]. 25. Gao, H., Struble, T. J., Coley, C. W., Wang, Y., Green, W. H., & Jensen, K. F. (2018). Using machine learning to predict suitable conditions for organic reactions. ACS Central Science, 4(11), 1465–1476. https://doi.org/10.1021/acscentsci.8b00357

[32]. [31] D. Weininger, "SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules," Journal of Chemical Information and Computer Sciences, vol. 28, no. 1, pp. 31–36, 1988, doi: 10.1021/ci00057a005.

[33]. J. P. Wolfe and S. L. Buchwald, "Asymmetric and regioselective intramolecular Heck reactions," Journal of the American Chemical

Society, vol. 119, no. 24, pp. 6054–6058, 1997, doi: 10.1021/ja970529l.

[34]. K. Sonogashira, "Development of Pd–Cu catalyzed cross-coupling of terminal acetylenes with sp2-carbon halides," Journal of Organometallic Chemistry, vol. 653, no. 1–2, pp. 46–49, 2002, doi: 10.1016/S0022-328X(02)01145-0.

[35]. A. F. Schmidt, A. A. Kurokhtina, and E. V. Larina, "The Suzuki–Miyaura reaction and the pharmaceutical industry," Current Organic Chemistry, vol. 14, no. 3, pp. 255–280, 2010, doi: 10.2174/138527210790231995.

[36]. Zhou, J. Recent advances in palladium-catalyzed Heck reactions of aryl halides with alkenes. ChemCatChem, 2011, 3(4), 513–528. doi: 10.1002/cctc.201000378

[37]. Dou, X., Feng, C., Loh, T.P. Recent advances in copper-catalyzed Sonogashira coupling reactions. Chemistry – A European Journal, 2017, 23(1), 8–14. doi: 10.1002/chem.201603662

[38]. González-Bobes, F., Fu, G.C. Kinetic and mechanistic studies of the palladium-catalyzed Suzuki cross-coupling of arylboronic acids and aryl bromides: Rate acceleration by hydroxide ions. Journal of the American Chemical Society, 2006, 128(16), 5360–5361. doi: 10.1021/ja060331o

[39]. Sayle, R.A., Tautomeric transformations using SMIRKS: Applications in chemical informatics. Journal of Chemical Information and Modeling, 2010, 50(12), 2236–2242. doi: 10.1021/ci100384d

Spectrum of Engineering Sciences
SPECTRUM OF ENGINEERING SCIENCES
Online ISSN
3007-3138
Print ISSN
3007-312X
SES

[40]. Schneider, N., Lowe, D.M., Sayle, R.A., Tarselli, M.A., Landrum, G.A. Big data from pharmaceutical patents: A computational analysis of medicinal chemists' bread and butter. Journal of Medicinal Chemistry, 2016, 59(9), 4385–4402. doi: 10.1021/acs.jmedchem.6b00109

[41]. Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., Kudlur, M., Levenberg, J., Monga, R., Moore, S., Murray, D.G., Steiner, B., Tucker, P., Vasudevan, V., Warden, P., Wicke, M., Yu, Y., Zheng, X. TensorFlow: A system for large-scale machine learning. Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 2016), 2016, pp. 265–283.

[42]. Read, J., Pfahringer, B., Holmes, G., & Frank, E. "Classifier chains for multi-label classification." *Machine Learning*, vol. 85, no. 3, pp. 333–359, 2011. doi: 10.1007/s10994-011-5256-5.

[43]. Gardner, M. W., & Dorling, S. R. "Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences." *Atmospheric Environment*, vol. 32, no. 14-15, pp. 2627–2636, 1998. doi: 10.1016/S1352-2310(97)00447-0.

[44]. Friedman, J. H. "Greedy function approximation: A gradient boosting machine." *Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001. doi: 10.1214/aos/1013203451.

[45]. Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., Kudlur, M., Levenberg, J., Monga, R., Moore, S., Murray, D. G., Steiner, B., Tucker, P., Vasudevan, V., Warden, P., Wicke, M., Yu, Y., & Zheng, X.

"TensorFlow: A system for large-scale machine learning." in *Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI)*, 2016, pp. 265–283.

[46]. O'Boyle, N. M., Banck, M., James, C. A., Morley, C., Vandermeersch, T., & Hutchison, G. R. "Open Babel: An open chemical toolbox." *Journal of Cheminformatics*, vol. 3, no. 1, pp. 33, 2011. doi: 10.1186/1758-2946-3-33.

[47]. Todeschini, R., & Consonni, V.*Molecular Descriptors for Chemoinformatics*. Wiley-VCH, 2009. doi: 10.1002/9783527628766.

[48]. Xia, X., Maliski, E. G., Gallant, P., & Rogers, D. "Classifying kinases using computational consensus sequence modeling." *Journal of Medicinal Chemistry*, vol. 47, no. 17, pp. 4463–4470, 2004. doi: 10.1021/jm0305862.

[49]. Manning, C. D., Raghavan, P., & Schütze, H.*Introduction to Information Retrieval*. Cambridge University Press, 2008.

[50]. Bergstra, J., Yamins, D., & Cox, D. D. "Making a Science of Model Search: Hyperparameter Optimization in Hundreds of Dimensions for Vision Architectures." in *Proc. 30th International Conference on Machine Learning (ICML 2013)*, 2013, pp. 115–123.

[51]. A. M. Zuranski and J. I. Martinez Alvarado, "Predicting reaction yields via supervised learning," *Accounts of Chemical Research*, vol. 54, no. 12, pp. 3172-3182, 2021. doi: 10.1021/acs.accounts.0c00770.

[52]. Md Tanvir Rahman Tarafder, Md Masudur Rahman, Nisher Ahmed, Tahmeed-Ur Rahman, Zakir Hossain, Asif Ahamed, "Integrating Transformative AI for Next-Level Predictive Analytics

in Healthcare," 2024 IEEE Conference on Engineering Informatics (ICEI-2024), Melbourne, Australia, 2024.

[53]. Asif Ahamed, Md Tanvir Rahman Tarafder, S M Tamim Hossain Rimon, Ekramul Hasan, Md Al Amin, "Optimizing Load Forecasting in Smart Grids with AI-Driven Solutions," 2024 IEEE International Conference on Data & Software Engineering (ICoDSE-2024), Gorontalo, Indonesia, 2024.

[54]. A. Nuthalapati, "Building Scalable Data Lakes For Internet Of Things (IoT) Data Management," Educational Administration: Theory and Practice, vol. 29, no. 1, pp. 412–424, Jan. 2023, doi: 10.53555/kuey.v29i1.7323.

[55]. SM T. H. Rimon, Mohammad A. Sufian, Zenith M. Guria, Niaz Morshed, Ahmed I. Mosaddeque, Asif Ahamed, "Impact of AI-Powered Business Intelligence on Smart City Policy-Making and Data-Driven Governance," International Conference on Green Energy, Computing and Intelligent Technology (GEn-CITy 2024), Johor, Malaysia, 2024.

[56]. Abbas, T., Khan, A. H., Kanwal, K., Daud, A., Irfan, M., Bukhari, A., & Alharbey, R. (2024). IoMT-Based Healthcare Systems: A Review. Computer Systems Science & Engineering, 48(4).

[57]. Abbas, T., Fatima, A., Shahzad, T., Alharbi, M., Khan, M. A., & Ahmed, A. (2024). Multidisciplinary cancer disease classification using adaptive FL in healthcare industry 5.0. Scientific Reports, 14(1), 18643.

[58]. Asif Ahamed, Nisher Ahmed, Jamshaid Iqbal Janjua, Zakir Hossain, Ekramul Hasan, Tahir Abbas, "Advances and Evaluation of

Intelligent Techniques in Short-Term Load Forecasting," 2024 International Conference on Computer and Applications (ICCA-2024), Cairo, Egypt, 2024.

[59]. T. Abbas, J. I. Janjua and M. Irfan, "Proposed Agricultural Internet of Things (AIoT) Based Intelligent System of Disease Forecaster for Agri-Domain," 2023 International Conference on Computer and Applications (ICCA), Cairo, Egypt, 2023, pp. 1-6, doi: 10.1109/ICCA59364.2023.10401794.

[60]. J. I., A. Sabir, T. Abbas, S. Q. Abbas and M. Saleem, "Predictive Analytics and Machine Learning for Electricity Consumption Resilience in Wholesale Power Markets," 2024 2nd International Conference on Cyber Resilience (ICCR), Dubai, United Arab Emirates, 2024, pp. 1-7, doi: 10.1109/ICCR61006.2024.10533004.

[61]. Nadeem, N., Hayat, M.F., Qureshi, M.A., et al., "Hybrid Blockchain-based Academic Credential Verification System (B-ACVS)," Multimed Tools Appl 82, 43991–44019, 2023. doi: 10.1007/s11042-023-14944-7.

[62]. A. Rehman, F. Noor, J. I. Janjua, A. Ihsan, A. Q. Saeed, and T. Abbas, "Classification of Lung Diseases Using Machine Learning Technique," 2024 International Conference on Decision Aid Sciences and Applications (DASA), Manama, Bahrain, 2024, pp. 1-7, doi: 10.1109/DASA63652.2024.10836302.

[63]. Asif Ahamed, Hasib Fardin, Ekramul Hasan, S M Tamim Hossain Rimon, Md Musa Haque, & Abdullah Al Sakib. (2022). Public Service Institutions Leading The Way With Innovative Clean Energy

Solutions . Journal of Population Therapeutics and Clinical Pharmacology, 29(04), 4477-4495.

[64]. J. I. Janjua, M. Nadeem and Z. A. Khan, "Machine Learning Based Prognostics Techniques for Power Equipment: Comparative Study," 2021 IEEE International Conference on Computing (ICOCO), Kuala Lumpur, Malaysia, 2021, pp. 265-270, doi: 10.1109/ICOCO53166.2021.9673564.

[65]. B. Y. Almansour, A. Y. Almansour, J. I. Janjua, M. Zahid, and T. Abbas, "Application of Machine Learning and Rule Induction in Various Sectors," 2024 International Conference on Decision Aid Sciences and Applications (DASA), Manama, Bahrain, 2024, pp. 1-8, doi: 10.1109/DASA63652.2024.10836265.

[66]. A. M. A. Al-Tarawneh, R. A. AlOmoush, T. ul Islam, J. I. Janjua, T. Abbas, and A. Ihsan, "Current Trends in Artificial Intelligence for Educational Advancements," 2024 International Conference on Decision Aid Sciences and Applications (DASA), Manama, Bahrain, 2024, pp. 1-6, doi: 10.1109/DASA63652.2024.10836340.

[67]. M. A. Sufian, S. M. T. H. Rimon, A. I. Mosaddeque, Z. M. Guria, N. Morshed, and A. Ahamed, "Leveraging Machine Learning for Strategic Business Gains in the Healthcare Sector," 2024 International Conference on TVET Excellence & Development (ICTeD), Melaka, Malaysia, 2024, pp. 225-230, doi: 10.1109/ICTeD62334.2024.10844658.

[68]. A. Nuthalapati, "Smart Fraud Detection Leveraging Machine Learning For Credit Card Security," Educational Administration:

Theory and Practice, vol. 29, no. 2, pp. 433–443, 2023, doi: 10.53555/kuey.v29i2.6907.

[69]. T. M. Ghazal, J. I. J, W. Abushiba, and S. Abbas, "Optimizing Patient Outcomes with AI and Predictive Analytics in Healthcare," 2024 IEEE 65th International Scientific Conference on Power and Electrical Engineering of Riga Technical University (RTUCON), Riga, Latvia, 2024, pp. 1-6, doi: 10.1109/RTUCON62997.2024.10830874.

[70]. S. B. Nuthalapati, M. Arun, C. Prajitha, S. Rinesh and K. M. Abubeker, "Computer Vision Assisted Deep Learning Enabled Gas Pipeline Leak Detection Framework," 2024 5th International Conference on Smart Electronics and Communication (ICOSEC), Trichy, India, 2024, pp. 950-957, doi: 10.1109/ICOSEC61587.2024.10722308.

[71]. J. I., S. Zulfiqar, T. A. Khan and S. A. Ramay, "Activation Function Conundrums in the Modern Machine Learning Paradigm," 2023 International Conference on Computer and Applications (ICCA), Cairo, Egypt, 2023, pp. 1-8, doi: 10.1109/ICCA59364.2023.10401760.

[72]. A. Nuthalapati, "Architecting Data Lake-Houses in the Cloud: Best Practices and Future Directions," Int. J. Sci. Res. Arch., vol. 12, no. 2, pp. 1902-1909, 2024, doi: 10.30574/ijsra.2024.12.2.1466.

[73]. Suri Babu Nuthalapati, "AI-Enhanced Detection and Mitigation of Cybersecurity Threats in Digital Banking," Educational Administration: Theory and Practice, vol. 29, no. 1, pp. 357–368, 2023, doi: 10.53555/kuey.v29i1.6908.

[74]. T. M. Ghazal et al., "Fuzzy-Based Weighted Federated Machine Learning Approach for Sustainable Energy Management with IoE Integration," 2024 Systems and Information Engineering Design Symposium (SIEDS), Charlottesville, VA, USA, 2024, pp. 112-117, doi: 10.1109/SIEDS61124.2024.10534747.

[75]. V. Voinarovska, M. Kabeshov, D. Dudenko, "When yield prediction does not yield prediction: an overview of the current challenges," ACS Journal of Chemical Information and Modeling, vol. 3, 2023, doi: 10.1021/acs.jcim.3c01524.

[76]. P. Schwaller, A.C. Vaucher, T. Laino, "Prediction of chemical reaction yields using deep learning," Machine Learning in Chemistry, vol. 7, 2021, doi: 10.1088/2632-2153/abc81d.

[77]. M.A. Kayala, P. Baldi, "ReactionPredictor: prediction of complex chemical reactions at the mechanistic level using machine learning," Journal of Chemical Information and Modeling, vol. 52, no. 3, pp. 563–571, 2012, doi: 10.1021/ci3003039.

[78]. G. Skoraczyński, P. Dittwald, B. Miasojedow, "Predicting the outcomes of organic reactions via machine learning: are current descriptors sufficient?" Scientific Reports, vol. 7, 2017, doi: 10.1038/s41598-017-02303-0.

[79]. J.M. Granda, L. Donina, V. Dragone, D.L. Long, L. Cronin, "Controlling an organic synthesis robot with machine learning to search for new reactivity," Nature, vol. 561, pp. 109–113, 2018, doi: 10.1038/s41586-018-0307-8.