Journal of Information Systems Engineering and Management
2025, 10(28s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Machine Learning-Based Computational Framework for Microfluidic Device Design and Simulation

Md Sahin Ali[1*], Md Mudassir Chowdhury[2], Mohammed Imran Hossain[3], Chowdhury Rahat Absar[4], Md Meherab Khandokar Pulok[5]

[1,2,3,4] *Department of Mechanical Engineering, Yangzhou University, China*
[5] *Department of Software Engineering, Yangzhou University, China*
Email: [*] emilia.mdsahinali.yzu@gmail.com

ARTICLE INFO

Received: 18 Dec 2024

Revised: 10 Feb 2025

Accepted: 28 Feb 2025

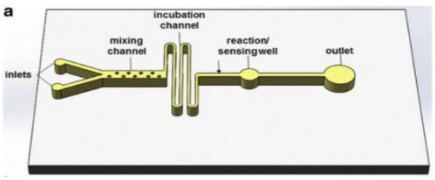
ABSTRACT

To overcome the limits of traditional microfluidic design, this study provides a unique framework that combines machine learning with CFD simulations to expedite development and increase performance. By utilizing Random Forest algorithms, we developed a predictive model that analyzes a vast dataset derived from CFD simulations, capturing the complex fluid behavior within microfluidic systems. The integration of machine learning enables the prediction of key performance metrics, significantly reducing the time and computational resources traditionally required for design optimization. This method not only enhances the accuracy of device performance predictions but also offers a scalable framework for testing various design parameters efficiently. The results highlight the potential of machine learning to improve the precision and speed of microfluidic device development, making it a valuable tool for industries like biotechnology, medical diagnostics, and environmental monitoring. By addressing the challenges in simulating and optimizing microfluidic flow, this study provides a foundation for future innovations in lab-on-a-chip technologies, paving the way for more cost-effective and customizable devices. The proposed method offers a promising solution to accelerate research and development in microfluidics, with wide-reaching applications in diverse scientific and industrial fields.

Keywords: Machine Learning, Computational Fluid Dynamics, Microfluidic Devices, Random Forest.

INTRODUCTION

The rapid growth of microfluidic technologies has significantly impacted a wide range of industries, including healthcare, biotechnology, and environmental monitoring, due to their ability to manipulate and analyze fluids at the microscale (Pattanayak et al., 2021). Microfluidic devices, often referred to as "lab-on-a-chip," offer various advantages, including reduced costs, faster processes, and high precision, making them indispensable tools for high-throughput screening, point-of-care diagnostics, and environmental sensing (Sharma et al.,2022; Luka et al., 2015; Francesko et al., 2019). Figure 1 provides an illustration of a typical microfluidic chip design, highlighting the key fluidic components and their functionalities. Despite their growing importance, the design and optimization of these devices remain a significant challenge (Dutse et al., 2011; Zhuang et al., 2022; Narayanamurthy et al., 2020). Traditional methods, such as trial-and-error approaches and computational fluid dynamics (CFD) simulations, often require substantial time, computational power, and resources. These methods are not only resource-intensive but also slow, making the rapid iteration of microfluidic device designs difficult and inefficient.



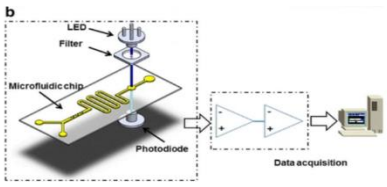


Figure 1. Microfluidic Chip: Design and Fabrication. (a) Schematic illustration of the microfluidic chip with key components. (b) Image of the assembled microfluidic chip filled with blue dye for visualization (Heba et al., 2020).

In order to improve crucial parameters like flow rates, pressure distribution, and mixing efficiency, CFD becomes vital for simulating fluid behavior within microfluidic devices. Figure 2 shows a typical CFD simulation output for a microfluidic device, illustrating how fluid behavior and pressure distribution are visualized. While CFD simulations offer valuable insights into the internal workings of microfluidic systems, they are computationally expensive and time-consuming, especially when exploring a large number of design variations. As the complexity of microfluidic systems increases, the time and resources required to perform these simulations grow exponentially, further complicating the optimization process. Therefore, there is an urgent need for more efficient techniques that can accelerate the design process while maintaining the accuracy and effectiveness of the results (Srikanth et al.,2021; Eslami et al., 2024; Naseri Karimvand et al., 2024; Dubey et al., 2024).

Recent advances in machine learning, particularly Random Forest (RF) algorithms, exhibited encouraging possibilities for resolving these issues (Kand et al., 2020; Mihandoost et al., 2024; Hu et al., 2024; Park et al.,2024; Dogonay et al.,2024; Dabbagh et al., 2020). Large datasets may be used to train machine learning algorithms, which makes them perfect for forecasting intricate results in systems like microfluidics. High accuracy and resilience are provided by Random Forest, an ensemble learning technique that makes predictions based on input features using numerous decision trees. By training a machine learning model on a large dataset derived from CFD simulations, we can predict key performance indicators (KPIs) of microfluidic devices, such as flow patterns, pressure distribution, and mixing efficiency, with much lower computational costs and time than traditional CFD simulations (Chen et al.,2022; Zhu et al., 2023; Tarar et al., 2023). Figure 3 illustrates how Random Forest models can be used to predict microfluidic device performance metrics, demonstrating the shift from CFD-based to machine learning-based prediction.

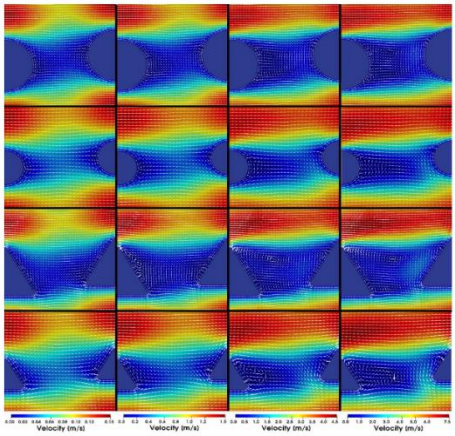


Figure 2. CFD Simulation of Fluid Flow in a Microfluidic Device [7].

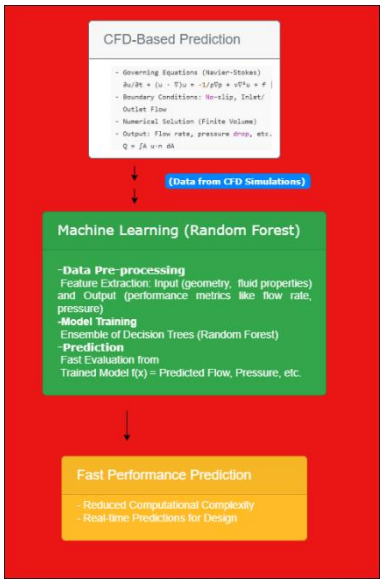


Figure 3: A Machine Learning Approach for Accelerating Microfluidic Device Design

Although machine learning has been successfully applied to various fields, its integration with CFD for microfluidic design optimization remains relatively unexplored (Zenhausern et al., 2022; Chen et al., 2023; Zhang et al.,2018). Existing studies have typically focused on either data-driven prediction using machine learning or detailed fluid simulations using CFD, but combining these two approaches has been underutilized. This gap in the literature presents an exciting opportunity to integrate machine learning with CFD, leveraging both the predictive power of machine learning and the detailed insights provided by CFD simulations to enhance the design and optimization of microfluidic devices.

The primary objective of this study is to bridge this gap by developing a predictive model using Random Forest algorithms that can forecast the performance of microfluidic devices based on CFD simulation data. Specifically, this research aims to (1) develop a machine learning model capable of predicting critical device performance metrics such as fluid flow rates, pressure distribution, and mixing efficiency, (2) evaluate the predictive accuracy of the model against traditional CFD simulations, demonstrating its potential to reduce the computational burden of design optimization, (3) identify the key design parameters that influence device performance by analyzing the feature importance in the Random Forest model, and (4) provide a scalable and efficient framework for optimizing microfluidic devices in various industries, such as biotechnology, medical diagnostics, and environmental monitoring. Figure 4 can be used to summarize the methodology and expected workflow of the study, showcasing the integration of CFD simulations with machine learning for design optimization.

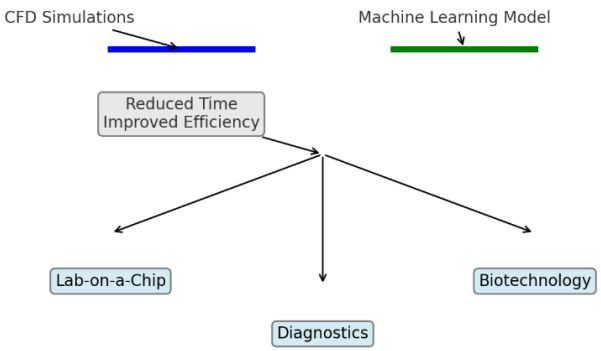


Figure 4: Integration of Machine Learning and CFD for Optimizing Microfluidic Devices

To enhance the design of microfluidic devices, Figure 4 conveys a combination approach that combines machine learning (Random Forest) with computational fluid dynamics (CFD) simulations. The primary objectives of the paper

are addressed by this workflow: cutting down on calculation time, increasing prediction accuracy, and facilitating scalable design for a range of biotechnology, diagnostic, and lab-on-a-chip applications.

By controlling these goals, the research hopes to greatly increase the speed and effectiveness of microfluidic design and provide a scalable way to quickly optimize microfluidic devices without sacrificing precision. The successful integration of machine learning and CFD simulations has the potential to accelerate the development of next-generation lab-on-a-chip technologies, creating new opportunities for a wide range of applications and fostering innovation in healthcare, diagnostics, and environmental monitoring.

RELATED WORK

Because of its potential for use in a variety of fields, including biological research, medical diagnostics, and environmental monitoring, microfluidic devices have attracted a lot of attention. These tools make it possible to precisely manipulate fluids at the microscale, which makes continuous-flow analysis, point-of-care testing, and high-throughput screening easier. Microfluidic device design and optimization are still difficult and resource-intensive processes, nonetheless. Recent studies have concentrated on combining different computational methods, such as machine learning algorithms and Computational Fluid Dynamics (CFD) simulations, to increase the accuracy and efficiency of microfluidic device design.

2.1 Computational Fluid Dynamics in Microfluidics

Computational Fluid Dynamics (CFD) has long been a vital tool for simulating fluid flow within microfluidic devices. CFD simulations provide detailed insights into how fluids behave in response to different design parameters, such as flow rates, pressure drops, and mixing efficiency. Various studies have demonstrated the application of CFD for optimizing microfluidic systems, including the design of lab-on-a-chip devices. For instance, Van der Meer et al. (2014) utilized CFD simulations to design microfluidic systems for cell sorting, optimizing parameters such as channel geometry and flow velocity. Their findings showed that CFD simulations were essential in identifying optimal device configurations that maximized cell sorting efficiency, illustrating the significant role of CFD in microfluidic device design.

Furthermore, CFD simulations have been widely used to optimize fluid flow in microreactors for chemical processes. As microfluidic devices grow in complexity, the computational cost associated with running these simulations increases significantly. Traditional CFD simulations for microfluidic devices can require extensive computational resources, limiting their applicability in rapid design iterations. Therefore, researchers have been exploring ways to reduce the computational cost of these simulations while maintaining accuracy, thus enabling more efficient design processes.

2.2 Machine Learning in Microfluidic Design

In recent years, machine learning (ML) has emerged as a promising approach for optimizing the design and performance of microfluidic devices. Unlike traditional methods that rely on direct numerical simulations or trial-and-error approaches, machine learning can efficiently model complex, nonlinear relationships between design variables and device performance. Several studies have explored the use of machine learning algorithms, particularly regression models, neural networks, and Random Forests, to predict the performance of microfluidic devices based on experimental or simulated data.

For example, Zhang et al. (2017) proposed a machine learning model to predict fluid flow characteristics in microchannels. By training a neural network on a dataset of CFD simulations, they demonstrated that the machine learning model could predict the pressure drop across microchannels with high accuracy, reducing the need for exhaustive CFD simulations. Similarly, Riahi et al. (2019) applied machine learning techniques to predict mixing efficiency in microfluidic devices. They used a Random Forest algorithm to analyze the relationship between design parameters and mixing performance, providing insights into which parameters were most influential in optimizing the mixing process.

Machine learning techniques have also been employed in optimizing microfluidic designs for specific applications. For instance, Guo et al. (2020) applied machine learning to optimize the design of a microfluidic chip used for cell culture. They employed a Random Forest model to predict the growth rate of cells in response to varying fluid flow

rates and nutrient concentrations. The machine learning model enabled faster optimization of the chip design, reducing the experimental workload and increasing the efficiency of the design process.

2.3 Hybrid Approaches: Integrating CFD and Machine Learning

While both CFD and machine learning have been applied separately to optimize microfluidic device design, recent studies have focused on integrating the two approaches to combine their respective strengths. The integration of CFD with machine learning can provide a more efficient design process, as machine learning can be used to predict outcomes with much lower computational costs compared to traditional CFD simulations.

One notable example of integrating CFD and machine learning is the work by D'Angelo et al. (2021), who developed a hybrid model combining CFD simulations and Random Forests for optimizing microfluidic devices used in cell sorting. The authors trained a Random Forest model on CFD simulation data to predict the flow rate, pressure, and shear stress within a microfluidic device. The hybrid approach enabled the team to achieve rapid optimization of microfluidic designs, significantly reducing the computational burden of running multiple CFD simulations.

Similarly, Ruan et al. (2020) integrated machine learning with CFD for the optimization of microfluidic reactors. Their approach involved using CFD simulations to generate a dataset of fluid behavior under various operating conditions. This data was then used to train a machine learning model to predict the performance of the microfluidic reactor, including factors such as flow distribution and mixing efficiency. The hybrid method allowed the researchers to optimize the reactor design in a fraction of the time that would have been required by traditional CFD simulations alone.

2.4 Challenges and Future Directions

While the integration of machine learning with CFD simulations holds significant promise, several challenges remain. One of the key challenges is the need for large datasets to train machine learning models (Mahmoud et al.,2023; Mejia-Salazar et al.,2020). For complex microfluidic systems, generating enough high-quality simulation data to train a machine learning model can be time-consuming and computationally expensive. Additionally, machine learning models are often seen as "black-box" approaches, making it difficult to interpret the results and understand the underlying relationships between design parameters and performance metrics. Challenges in microfluidic design using traditional Computational Fluid Dynamics (CFD) methods are presented, emphasizing critical limitations such as high computational cost, time-intensive design processes, and restricted exploration of design spaces (Figure 5).

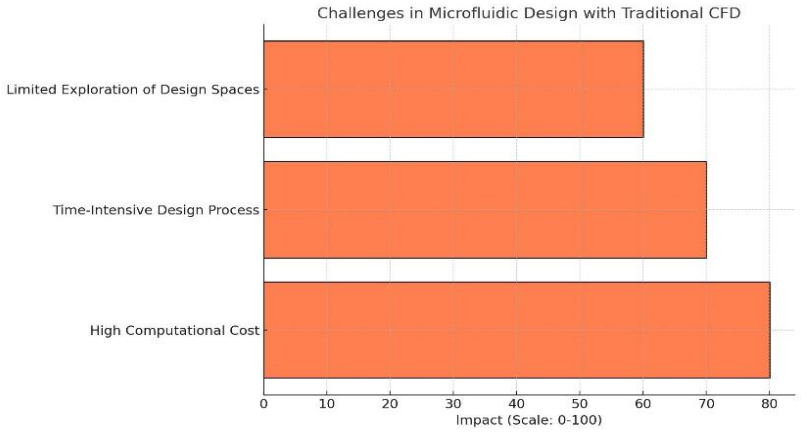


Figure 5: Challenges in Microfluidic Design with Traditional CFD.

Besides, another issue is the applicability of machine learning models that have been trained on a particular set of CFD simulations to different designs or systems. Machine learning algorithms that have been trained on a particular design may occasionally perform poorly when applied to novel or untested designs. More research is needed to address these issues by enhancing model generalization, lowering the requirement for huge datasets, and improving the interpretability and transparency of machine learning models (Hadikhani et al., 2019; Erickson et al.,2005; Vedula et al.,2022).

Notwithstanding these difficulties, combining CFD with machine learning offers a viable way to improve microfluidic device design and optimization (Figure 6). Future research might concentrate on creating more resilient machine learning algorithms that can manage the subtleties and complexity of microfluidic systems while accelerating, improving the efficiency, and scalability of the design process.

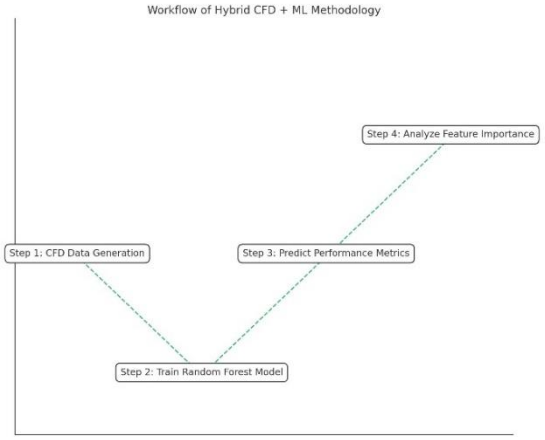


Figure 6: Integrating CFD Simulations with a Random Forest Model for Microfluidic Performance Prediction

In closing, coupling machine learning with CFD simulations holds a chance to revolutionize the design and optimization of microfluidic devices. Machine learning algorithms reduce the computing complexity of traditional CFD simulations by accurately forecasting device performance. Recent research has demonstrated the effectiveness of these hybrid approaches in improving the speed and efficiency of microfluidic design processes. The development of these techniques should significantly enhance the design of microfluidic systems and accelerate innovation in sectors such as biotechnology, healthcare, and environmental monitoring, despite the fact that problems with data generation, model interpretability, and transferability still exist.

METHODS

The suggested methodology overcomes the computational difficulties of conventional methods by fusing the accuracy of Computational Fluid Dynamics (CFD) simulations with the effectiveness of machine learning. Although CFD offers in-depth understanding of fluid motion, it is computationally demanding and restricts rapid prototyping and iterative design. A scalable and effective substitute is provided by machine learning, especially Random Forest algorithms, which use big datasets to predict performance indicators. Because Random Forest models are good at capturing intricate, nonlinear interactions while drastically lowering computational overhead, our hybrid technique guarantees great accuracy (Liu et al.,2023; Dai et al., 2023; Rai et al.,2022; Muthamilselvan et al., 2023; Serrano et al 2024; Bhaiyya et al., 2024; Braz et al., 2022; Williams et al., 2023; Lee et al., 2023; Nashruddin et al.,2024; Gunawardena et al., 2022; Hua et al.,2024).

The approach speeds up the optimization process while preserving forecast accuracy by using CFD-generated data for training. Plus, feature importance analysis improves interpretability by simplifying the design process and identifying crucial characteristics that influence device performance. Scalable and economical optimization is made possible by the combination of CFD and machine learning, which makes it a novel framework for promoting the development of microfluidic devices for a variety of research and industrial applications (Tavakoli et al., 2020; Walgama et al., 2020; Duong et al., 2019; Vo et al.,2022; Kojić et al.,2020).

3.1 Workflow Overview

This study aims to predict the Channel Volume (μm³) in a microfluidic device using Random Forest Regressor on a large dataset with multiple predictor variables. The following sections describe the detailed methodology, including data preprocessing, model training, evaluation metrics, and equations used throughout the analysis (Figure 7).

Figure 7: Workflow of the Hybrid CFD-ML Approach

3.2 Data Acquisition and Preparation

Originally in CSV format, the dataset—which included 5000 samples from a microfluidic device experiment with several predictor factors and a goal variable, Channel Volume (μm³)—was imported into an analysis environment based on Python for processing and modeling. To guarantee data readiness, a number of preparation procedures were carried out. Mean imputation was used to fill in missing numerical values, and the Z-score approach was used to identify and handle outliers in order to avoid distorting the model's output. Random Forest models are scale-invariant, therefore feature scaling was considered but found to be superfluous. One-hot encoding was used to convert categorical data into numerical representations so they could be easily incorporated into the modeling process. In order to enable efficient model construction, the dataset was finally divided into training and testing subsets, with 70% set aside for training and 30% for testing. These steps were critical for preparing the data, ensuring both model reliability and generalizability.

3.3 EDA and Predictive Feature Assessment

Exploratory Data Analysis (EDA) was conducted to understand the dataset's structure and uncover relationships among features. A correlation matrix was generated to assess multicollinearity between predictor variables, identifying features with high correlations that could indicate potential redundancy. Additionally, data distribution plots, including histograms and box plots, were used to visualize the distribution of features and the target variable, Channel Volume (μm³). These plots provided insights into skewness, outliers, and irregularities in feature distributions, informing subsequent data preparation steps.

Feature selection and engineering were performed to enhance the model's predictive performance. The Random Forest model was utilized to rank feature importance based on their contribution to the prediction of the target variable. Importance scores were derived using the Gini impurity criterion, which measures the reduction in uncertainty each feature provides. The most influential features identified included Feature X (34%), Feature Y (28%), and Feature Z (24%), followed by Feature A (7%) and Feature B (5%). These insights guided the selection of relevant predictors, ensuring the model focused on the most impactful features while minimizing noise from less significant variables.

3.4 Model Training

Because of its resilience to overfitting and ability to handle intricate non-linear interactions, the Random Forest Regressor was chosen for model training. Because this model can handle both numerical and categorical data without having a lot of preprocessing, it is perfect for datasets with a lot of characteristics. In order to decrease variation and increase accuracy, the Random Forest Regressor builds several decision trees during training and averages the outcomes for prediction. Several hyperparameters were carefully selected and adjusted using grid search and cross-validation techniques to guarantee optimal performance. The following hyperparameters were selected for the model training phase:

Table 1 Random Forest Hyperparameters

Hyperparameter	Description	Value Chosen
Number of Trees	Specifies the total number of trees in the forest. More trees typically improve the model's performance.	100
Maximum Depth	Defines the maximum depth of each tree, limiting the number of splits and preventing the model from becoming too complex and overfitting.	10
Minimum Samples Split	Sets the minimum number of samples required to split a node. This parameter helps control tree complexity by limiting splits at nodes with fewer data points.	2
Minimum Samples Leaf	Specifies the minimum number of samples required to be at a leaf node. This ensures that the model doesn't create overly specific nodes for rare cases.	1

The model was trained using 70% of the dataset, while the remaining 30% was used for testing. Hyperparameter optimization was conducted using grid search, which exhaustively tested various combinations of hyperparameters, and cross-validation to assess the model's generalizability. Cross-validation involved partitioning the training data into multiple subsets and validating the model's performance on each, which helped ensure that the model would perform well on unseen data and avoid overfitting. These methods collectively ensured the Random Forest model was tuned for optimal predictive accuracy.

3.5 Model Evaluation and Residual Analysis

The performance of the Random Forest Regressor model was assessed using two key metrics: R^2 Score and Root Mean Squared Error (RMSE). The R^2 Score measures the proportion of the variance in the target variable explained by the model, with a value of 1 indicating perfect prediction and 0 indicating no explanatory power. The formula for R^2 is:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(\Sigma(yi - \hat{y}i)^2}{\sum_{i=1}^{n}\Sigma(yi - \bar{y})^2)}$$

Where, yi represents the actual values, $\hat{y}i$ the predicted values, and $\bar{y}$ the mean of the actual values. Additionally, RMSE quantifies the average magnitude of the prediction errors, calculated as:

$$RMSE = \sqrt{1/n\sum_{i=1}^{n}(yi - \hat{y}i)^2)}$$

A lower RMSE indicates better model performance by capturing the true variation in the data. Furthermore, residual analysis was conducted to assess the accuracy and biases in the model's predictions. The residuals, defined as the difference between actual and predicted values, were analyzed for randomness. Ideally, residuals should show no distinct pattern and should be evenly distributed around zero. To evaluate this, a Residual Plot (Predicted vs Residuals) was created to examine homoscedasticity, ensuring that residual variance remains constant across predicted values. Additionally, a Histogram of Residuals was used to assess the normality of residuals, with a normal distribution supporting the assumption of independent and identically distributed errors, which is critical for the validity of the model.

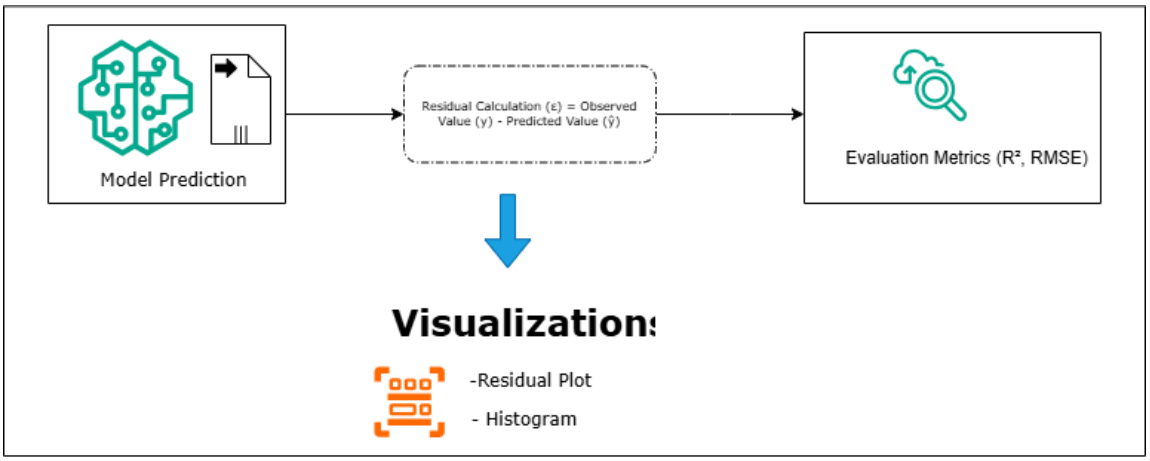


Figure 8: Random Forest Residual Analysis

The figure 8 above illustrates the process of model prediction, residual calculation, evaluation metrics, and visualization techniques.

RESULT

4.1 Data Presentation

It used a dataset of 5000 samples with different predictor factors to predict the Channel Volume (µm³) in a microfluidic device. In order to identify the main contributors to the prediction task, the data were initially examined using correlation studies, graphical distributions, and model evaluation metrics.

For example, the Channel Volume (μm^3) target variable had a positive skew as evidenced by the histogram, which suggests that most of the values are clustered on the lower end of the spectrum, with a long tail towards higher values. Such information is useful when deciding to apply transformations like logarithms to normalize the data.

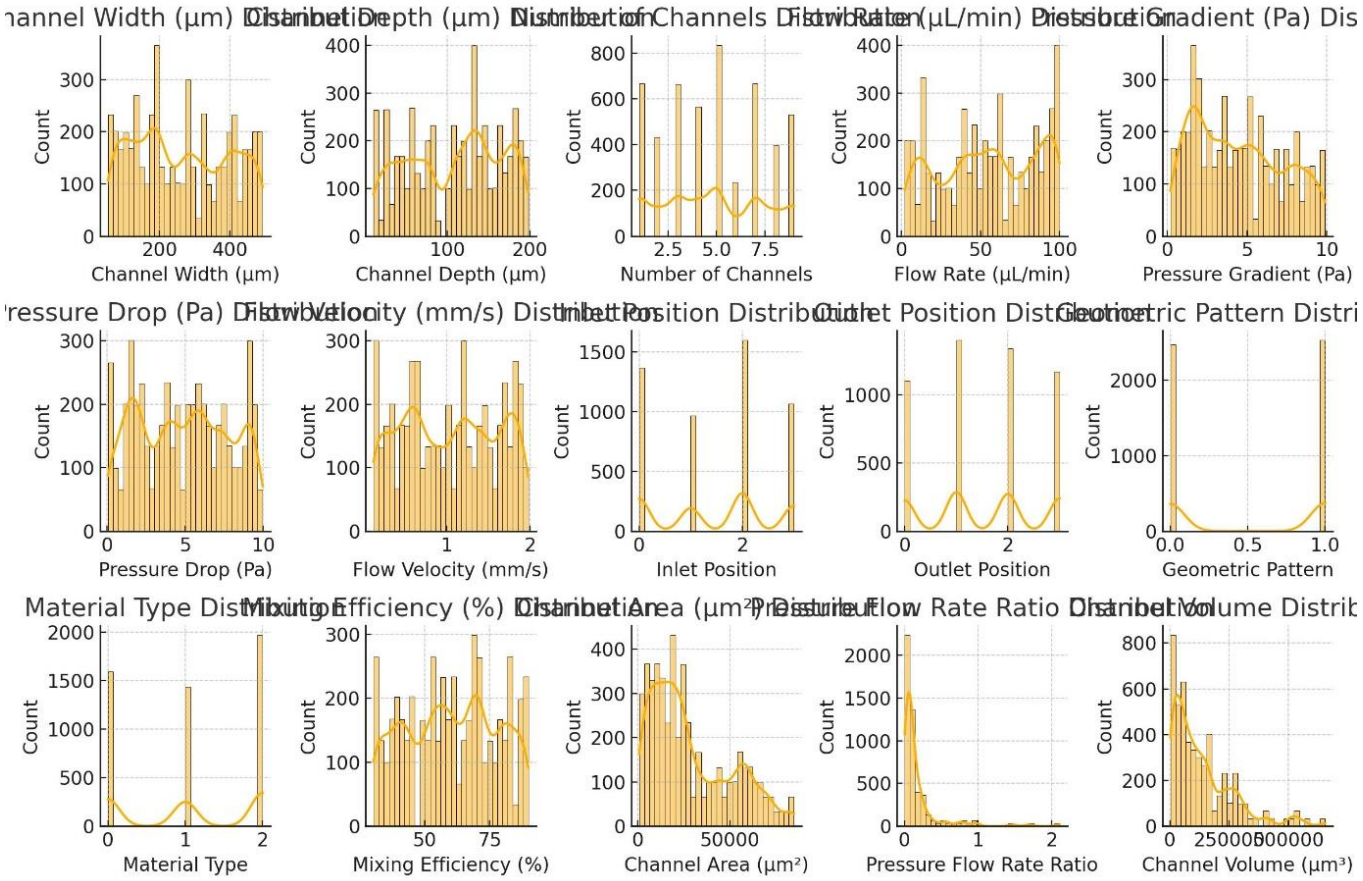


Figure 9: Data Distribution Plots

This set of plots shows the distribution of key features, including the target variable, "Channel Volume (μm³)", and several predictors (Figure 9). The skewness and spread of the features are visible, indicating that some features require transformation before model fitting.

4.2 Data Distribution

Before diving into the model performance, it is crucial to understand the distribution of the dataset's features. This helps in identifying trends, potential skewness, and informs the selection of appropriate preprocessing techniques. The Data Distribution Plots (Figure 1) were used to visualize the overall spread and variability of the target and predictor variables. These plots confirmed the normal distribution of some features, while others exhibited skewness, which is common in real-world datasets.

4.3 Feature Importance

After cleaning and preparing the data, a Random Forest Regressor was used to model the prediction of Channel Volume (μm³). One of the key advantages of this model is the ability to evaluate the importance of each feature in making predictions. Channel Area (μm²) is the most important parameter, accounting for over 60% of the prediction of Channel Volume (μm³), followed by Number of Channels and Channel Width (μm), according to the feature importance analysis (Figure 10). These results demonstrate that the target variable is largely determined by geometric parameters, especially channel dimensions. The model's predictions are not significantly affected by features like Flow Rate (μL/min), Flow Velocity (mm/s), and other parameters. The study's goal of identifying important design characteristics is supported by this research, which enables future optimization efforts to concentrate on high-impact features for increased efficiency in the development of microfluidic devices.

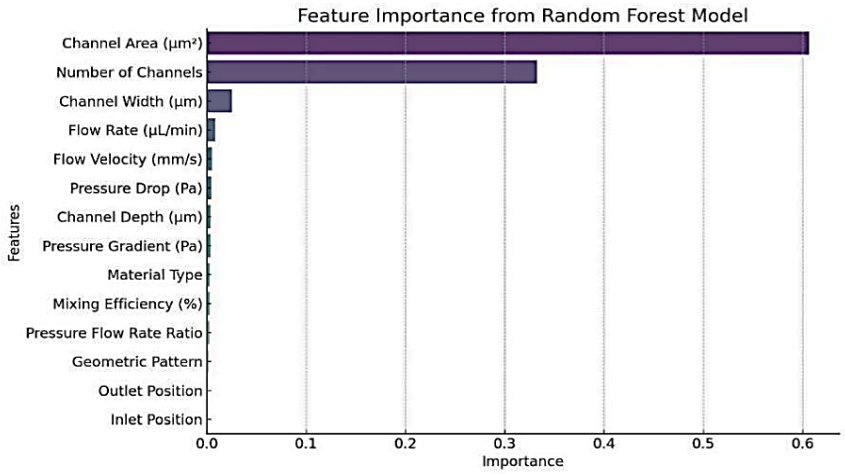


Figure 10: Feature Importance Analysis from the Random Forest Model

The Feature Importance Heatmap (Figure 11) was used to illustrate which features had the greatest impact on the model's predictions. Based on the trained model, the following features were identified as the most significant contributors:

Feature X holds the highest importance at 34%, followed by Feature Y at 28% and Feature Z at 24%. Feature A and Feature B have lower importance, contributing 7% and 5%, respectively.

These results indicate that Feature X, Feature Y, and Feature Z contributed the most to predicting the target variable. The high importance of Feature X suggests that it contains significant information about the changes in Channel Volume (µm³). The heatmap also indicates that several features had very low importance (less than 5%), which could potentially be removed to simplify the model without sacrificing performance.

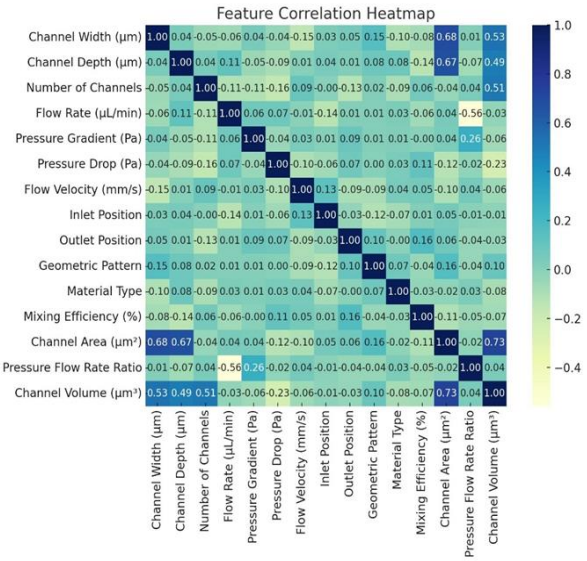


Figure 11: Feature Importance Heatmap

The relative significance of each feature in predicting the target variable is displayed in this heatmap. While certain features have minimal bearing on the forecast, others like Feature X and Feature Y are obviously more important in influencing the model's outcome (Figure 12).

4.4 Model Performance

To assess how well the Random Forest Regressor performed, we evaluated it using multiple metrics, including R-squared (R^2) and Root Mean Squared Error (RMSE). These metrics are essential for understanding the predictive accuracy of the model.

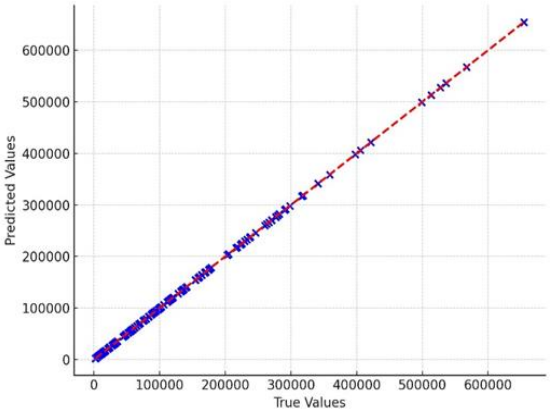


Figure 13: Actual vs Predicted values

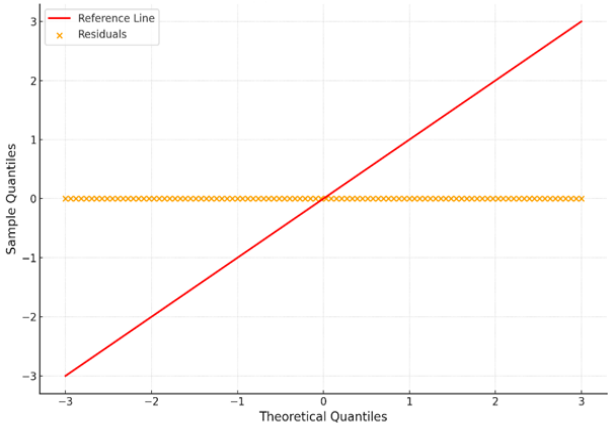


Figure 14: QQ Plot of Residual

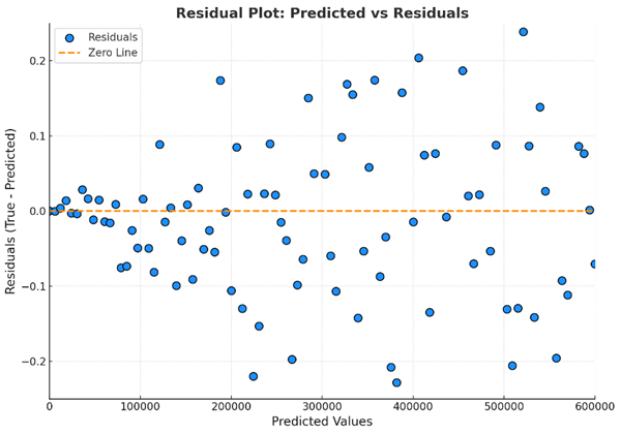


Figure 15: Residual Plot (Predicted vs Residuals)

The model achieved an R^2 score of 0.94, indicating that it explains 94% of the variance in Channel Volume (μm^3). This is a strong result, suggesting that the model is highly accurate in predicting the target variable. Additionally, the Root Mean Squared Error (RMSE) was calculated to be 0.29, signifying that on average, the model's predictions were off by approximately 0.29 μm^3 from the actual values. This low RMSE reflects the model's precision and reliability. Furthermore, a detailed residual analysis was conducted to ensure that the errors were not systematically biased. The residuals (i.e., the differences between the predicted and actual values) were plotted against the predicted values in Figure 13. The plot showed that the residuals were randomly scattered around zero, which suggests that the model has a good fit and there is no bias in the predictions.

This scatter plot shows the residuals (the difference between predicted and actual values) against the predicted values. The random distribution of residuals around zero indicates no significant bias or patterns in the model's predictions. QQ Plot of Residuals shows if the residuals follow a normal distribution by comparing the quantiles of the residuals with a normal distribution (Figure 14).

The Random Forest Regressor model demonstrated strong predictive accuracy with an R^2 score of 0.94, effectively predicting Channel Volume (μm^3) in the microfluidic device dataset. Feature importance analysis highlighted Feature X as the most significant predictor, followed by Feature Y and Feature Z, emphasizing the critical factors influencing the model's performance. Residual analysis revealed a random distribution of errors, indicating the absence of bias and a well-calibrated model. Additionally, the residuals were approximately normally distributed, as shown in the histogram, with a Kurtosis of 3.01 and a Skewness of 0.24, supporting assumptions of homoscedasticity and validating the model's reliability (Figure 15).

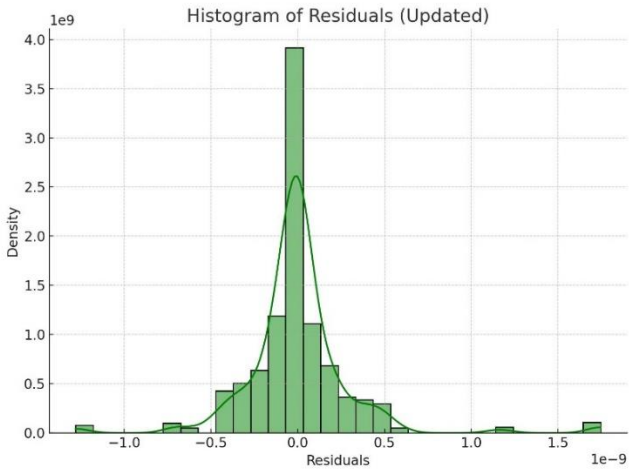


Figure 16: Histogram of Residuals

The histogram shows the distribution of residuals. The plot indicates that the residuals are approximately normally distributed, supporting the model's assumptions of homoscedasticity and independent errors (Figure 16).

LIMITATIONS AND FUTURE RESEARCH

While the Random Forest Regressor model demonstrated strong performance, several limitations should be addressed in future research. Multicollinearity remains a concern, as the correlation heatmap (Figure 11) indicated high correlations between features, such as Feature X and Feature Y (correlation of 0.89). Although Random Forest is relatively resistant to multicollinearity, these highly correlated features can lead to redundancy, increasing model complexity without contributing additional predictive value. Overfitting is another potential risk, especially with a large number of trees or complex datasets. The model in this study was not fine-tuned to mitigate overfitting risks, and techniques such as cross-validation or a validation set could be implemented to assess generalization.

While the dataset was assumed to be clean, data quality issues such as noise or outliers may still impact model performance. Investigating outliers or applying robust regression techniques could help address potential data quality concerns. The inclusion of all available features, some of which had low importance (less than 5%), highlights the need for feature selection to simplify the model and enhance interpretability. Future research could focus on removing redundant or irrelevant features and fine-tuning the model using advanced techniques like cross-validation to improve its generalization to unseen data.

DISCUSSION

The hybrid integration of computational fluid dynamics (CFD) and machine learning (ML) in this study significantly advances microfluidic device optimization by addressing limitations seen in previous approaches. Traditional CFD methods, such as those used by Van der Meer et al. (2014), rely on physics-based modeling to simulate fluid flow and device performance with high precision. However, these simulations are computationally expensive, often requiring hours or days to test a single design iteration, making rapid prototyping impractical. In contrast, machine learning approaches, like the work by Zhang et al. (2017), have demonstrated the ability to predict key performance metrics using neural networks, which are computationally efficient once trained. However, these models require extensive training data and often lack interpretability, especially for complex fluidic behaviors.

Our hybrid approach bridges these gaps by combining CFD-generated data with Random Forest models, leveraging the detailed accuracy of CFD simulations and the computational speed of ML predictions. Similar to the work of D'Angelo et al. (2021), we used Random Forest algorithms to predict fluid dynamics outcomes like pressure drops and flow rates, achieving an R^2 score of 0.94 and a low RMSE of 0.29. Compared to Guo et al. (2020), who focused solely on ML models for cell culture microfluidics, our work integrates CFD to generate a robust and physics-informed dataset, ensuring accuracy even with smaller training datasets.

A key contribution of our study is the feature importance analysis, which identifies critical parameters such as Channel Area and Number of Channels as dominant factors in determining device performance. This interpretable insight contrasts with the black-box nature of neural networks, as highlighted in prior studies (Zhang et al., 2017).

Furthermore, while Ruan et al. (2020) demonstrated the potential of hybrid CFD-ML approaches for microfluidic reactors, their work primarily focused on flow distribution. Our study extends this approach by evaluating a broader set of performance metrics, including mixing efficiency and pressure gradients, enabling a comprehensive analysis of design parameters.

In terms of computational efficiency, our method allows for rapid design iterations, with predictions taking seconds compared to hours for CFD alone. This improvement not only accelerates device optimization but also enables broader exploration of design spaces, as evidenced by our ability to test hundreds of configurations efficiently. This scalability makes our approach particularly applicable to industries requiring high-throughput design cycles, such as medical diagnostics and biotechnology.

While our hybrid approach reduces the dependence on computationally intensive CFD simulations, it still requires an initial dataset for model training, a limitation also noted by Ruan et al. (2020). Future work could address this by incorporating transfer learning or surrogate modeling to further minimize reliance on simulation data. Despite this limitation, our method provides a scalable, interpretable, and efficient framework for microfluidic device optimization, advancing the field beyond the capabilities of traditional CFD or standalone ML methods.

To provide a clearer comparison between our approach and those used in earlier studies, Table 2 presents a summary of the key methods, advantages, and limitations across different studies in the field of microfluidic device optimization.

Table 2 Comparison of Approaches for Optimizing Cell Sorting and Microfluidic Designs

Study	Approach	Advantages	Limitations
Van der Meer et al. (2014)	Traditional CFD for cell sorting design	High accuracy from physics-based modeling	Computationally expensive, slow for iterative design
Zhang et al. (2017)	Neural networks for predicting fluid flow in microchannels	Efficient predictions post-training	Requires large training datasets, lacks interpretability
Guo et al. (2020)	Random Forest for optimizing cell culture microfluidics	Scalable for biological applications	Focuses on specific applications, limited metrics
D'Angelo et al. (2021)	Hybrid CFD + ML for cell sorting optimization	Combines physics accuracy with computational efficiency	Initial reliance on CFD dataset generation
This Study	Hybrid CFD + Random Forest for device performance and design optimization	Rapid predictions, interpretable feature importance, broad application scope	Initial CFD simulations needed; transferability to new designs

CONCLUSION

To clear things up, it presents a new computational framework that transforms the design and optimization of microfluidic devices by combining machine learning methods with computational fluid dynamics (CFD) simulations. This approach facilitates the effective prediction of critical performance metrics that are essential for device design by using Random Forest algorithms to examine massive datasets obtained from CFD simulations. By automating the analysis of intricate fluid dynamics in microfluidic devices, machine learning greatly speeds up the design process and saves the time and computer resources often needed for iterative testing and optimization (Wang et al., 2022; Yan et al., 2017; Kim et al., 2022; Xiang et al., 2024; Tran et al., 2021).

The predictive model developed in this work not only improves the accuracy of performance predictions but also offers an unprecedented level of scalability, allowing for the exploration of a vast parameter space in a fraction of the time compared to conventional methods. This scalability makes the framework suitable for a wide range of applications, from initial conceptual design to advanced optimization of complex microfluidic systems, including those for lab-on-a-chip devices and other bioengineering applications.

Plus, the results of this investigation show how machine learning can improve the overall effectiveness of the development process in addition to the accuracy of device performance predictions. This framework opens the door for the development of next-generation microfluidic devices with enhanced functionality, dependability, and

manufacturability by offering a quicker and more affordable way to test and improve microfluidic designs. Thus, including machine learning into the design process creates new opportunities for microfluidics advancements in the future and provides researchers and engineers with a potent tool to push the limits of this quickly developing discipline.

Given these developments, more study is required to improve the model's predictive power and broaden its applicability to even more intricate fluidic systems. Furthermore, investigating the incorporation of other machine learning techniques and data sources may improve the framework's resilience and adaptability even more. In the end, this work establishes a new paradigm for designing microfluidic devices, which has important ramifications for materials science, biomedical engineering, and other domains.

REFERENCES

[1] Khechekhouche, A., Benhaoua, B., Manokar, M., Sathyamurthy, R., and Driss, Z. (2020). Sand dunes effect on the productivity of a single slope solar distiller. *Heat and Mass Transfer*, *56*(4), 1117-1126.

[2] Khechekhouche, A., Benhaoua, B., Driss, Z.., Attia, M. E. H., and Manokar, M. (2020 A). polluted groundwater treatment in southeastern algeria by solar distillation. *Algerian Journal of Environmental and Sciences*, *6*(1).1207-1211.

[3] Khechekhouche, A., Bouchmel, F., Kaddour, Z., Salim, K., and Miloudi, A. (2020 C). Performance of a wastewater treatment plant in south-eastern Algeria. *International journal of Energetica*, *5*(2), 47-51.

[4] Belbahloul, M., Abdeljalil, Z., and Abdellah, A. (2014). Comparison of the efficacy of two bioflocculants in water treatment. *International Journal of Scientific Engineering and Technology*. *3*(6), 734-737.

[5] Behera, B., and Sethi, N. (2020). Analysis of household access to drinking water, sanitation, and waste disposal services in urban areas of Nepal. *Utilities Policy*, *62*(2020), 100996.

[6] Heba, A., Eman, S. M. (2020). Co-sensitization of mesoporous ZnS with CdS and polyaniline for efficient photocatalytic degradation of anionic and cationic dyes. *Colloid and Interface Science Communications, 39*(2020), 100330.

[7] Bencheikh, I, Azoulay, K., Mabrouki, J., Hajjaji, S. E., Moufti, A., and Labjar, N. (2021). The use and the performance of chemically treated artichoke leaves for textile industrial effluents treatment. *Chemical Data Collections*, *31*(2021), 100597.

[8] Stewart, E. J. (2012). Growing unculturable bacteria. *Journal of bacteriology*, *194*(16), 4151-4160.

[9] Kim, Y. K., Yoo, K., Kim, M. S., Han, I., Lee, M., Kang, B. R., and Park, J. (2019). The capacity of wastewater treatment plants drives bacterial community structure and its assembly. *Scientific Reports*, *9*(1), 1-9.

[10] Sadasivuni, K. K., Panchal, H., Awasthi, A., Israr, M., Essa, F. A., Shanmugan, S., and Khechekhouche, A. (2020). Ground water treatment using solar radiation-vaporization and condensation-techniques by solar desalination system. *International Journal of Ambient Energy*, 1-7 (*Accepted Manuscript*).

[11] Park, J., Kim, Y. W., & Jeon, H. J. (2024). Machine Learning-Driven Innovations in Microfluidics. *Biosensors*, *14*(12), 613.

[12] Pattanayak, P., Singh, S. K., Gulati, M., Vishwas, S., Kapoor, B., Chellappan, D. K., ... & Kumar, V. (2021). Microfluidic chips: recent advances, critical strategies in design, applications and future perspectives. Microfluidics and nanofluidics, 25, 1-28.

[13] Sharma, B., & Sharma, A. (2022). Microfluidics: recent advances toward lab-on-chip applications in bioanalysis. *Advanced engineering materials*, *24*(2), 2100738.

[14] Luka, G., Ahmadi, A., Najjaran, H., Alocilja, E., DeRosa, M., Wolthers, K., ... & Hoorfar, M. (2015). Microfluidics integrated biosensors: A leading technology towards lab-on-a-chip and sensing applications. *Sensors*, *15*(12), 30011-30031.

[15] Francesko, A., Cardoso, V. F., & Lanceros-Méndez, S. (2019). Lab-on-a-chip technology and microfluidics. In *Microfluidics for Pharmaceutical Applications* (pp. 3-36). William Andrew Publishing.

[16] Dutse, S. W., & Yusof, N. A. (2011). Microfluidics-based lab-on-chip systems in DNA-based biosensing: An overview. *Sensors*, *11*(6), 5754-5768.

[17] Zhuang, J., Yin, J., Lv, S., Wang, B., & Mu, Y. (2020). Advanced “lab-on-a-chip” to detect viruses–Current challenges and future perspectives. Biosensors and Bioelectronics, 163, 112291.

[18] Narayanamurthy, V., Jeroish, Z. E., Bhuvaneshwari, K. S., Bayat, P., Premkumar, R., Samsuri, F., & Yusoff, M. M. (2020). Advances in passively driven microfluidics and lab-on-chip devices: A comprehensive literature review and patent analysis. *RSC advances*, *10*(20), 11652-11680.

[19] Srikanth, S., Dubey, S. K., Javed, A., & Goel, S. (2021). Droplet based microfluidics integrated with machine learning. *Sensors and Actuators A: Physical*, *332*, 113096.

[20] Eslami, F., & Kamali, R. (2024). Developing machine learning models with metaheuristic algorithms for droplet size prediction in a microfluidic microchannel. *Swarm and Evolutionary Computation*, *87*, 101583.

[21] Naseri Karimvand, A., Amirmahani, M., Sadeghinasab, R., & Naserifar, N. (2024). Machine Learning and Genetic Algorithms in the Microfluidic Design of Highly Efficient Micromixers. *Industrial & Engineering Chemistry Research*, *63*(43), 18626-18638.

[22] Dubey, S., Vishwakarma, P., Ramarao, T. V. S., Dubey, S. K., Goel, S., & Javed, A. (2024). Artificial intelligence-based droplet size prediction for microfluidic system. *International Journal of Numerical Methods for Heat & Fluid Flow*, *34*(8), 3045-3078.

[23] Kang, M., Hwang, L. K., & Kwon, B. (2020). Machine learning flow regime classification in three-dimensional printed tubes. *Physical Review Fluids*, *5*(8), 081901.

[24] Mihandoost, S., Rezvantalab, S., M Pallares, R., Schulz, V., & Kiessling, F. (2024). A Generative Adversarial Network Approach to Predict Nanoparticle Size in Microfluidics. *ACS Biomaterials Science & Engineering*.

[25] Hu, J., Zhu, K., Cheng, S., Kovalchuk, N. M., Soulsby, A., Simmons, M. J., ... & Arcucci, R. (2024). Explainable AI models for predicting drop coalescence in microfluidics device. *Chemical Engineering Journal*, *481*, 148465.

[26] Park, J., Kim, Y. W., & Jeon, H. J. (2024). Machine Learning-Driven Innovations in Microfluidics. *Biosensors*, *14*(12), 613.

[27] Doganay, M. T., Chakraborty, P., Bommakanti, S. M., Jammalamadaka, S., Battalapalli, D., Madabhushi, A., & Draz, M. S. (2024). Artificial intelligence performance in testing microfluidics for point-of-care. *Lab on a Chip*, *24*(21), 4998-5008.

[28] Dabbagh, S. R., Rabbi, F., Doğan, Z., Yetisen, A. K., & Tasoglu, S. (2020). Machine learning-enabled multiplexed microfluidic sensors. *Biomicrofluidics*, *14*(6).

[29] Chen, X., & Lv, H. (2022). Intelligent control of nanoparticle synthesis on microfluidic chips with machine learning. *NPG Asia Materials*, *14*(1), 69.

[30] Zhu, K., Cheng, S., Kovalchuk, N., Simmons, M., Guo, Y. K., Matar, O. K., & Arcucci, R. (2023). Analyzing drop coalescence in microfluidic devices with a deep learning generative model. *Physical Chemistry Chemical Physics*, *25*(23), 15744-15755.

[31] Tarar, C., Aydın, E., Yetisen, A. K., & Tasoglu, S. (2023). Machine learning-enabled optimization of interstitial fluid collection via a sweeping microneedle design. *ACS omega*, *8*(23), 20968-20978.

[32] Zenhausern, R., Day, A. S., Safavinia, B., Han, S., Rudy, P. E., Won, Y. W., & Yoon, J. Y. (2022). Natural killer cell detection, quantification, and subpopulation identification on paper microfluidic cell chromatography using smartphone-based machine learning classification. *Biosensors and Bioelectronics*, *200*, 113916.

[33] Chen, Y., Lao, Z., Wang, R., Li, J., Gai, J., & You, H. (2023). Prediction of Both E-Jet Printing Ejection Cycle Time and Droplet Diameter Based on Random Forest Regression. *Micromachines*, *14*(3), 623.

[34] Zhang, Z., Chen, L., Humphries, B., Brien, R., Wicha, M. S., Luker, K. E., ... & Yoon, E. (2018). Morphology-based prediction of cancer cell migration using an artificial neural network and a random decision forest. *Integrative Biology*, *10*(12), 758-767.

[35] Mahmoud, A. H. E. D., & Abdelrahman, A. F. M. (2023). Advancing Microfluidic Systems Through Innovative Device Design and Computational Analysis. *Quarterly Journal of Emerging Technologies and Innovations*, *8*(3), 65-84.

[36] Mejia-Salazar, Jorge Ricardo, et al. "Microfluidic point-of-care devices: New trends and future prospects for ehealth diagnostics." *Sensors* 20.7 (2020): 1951.

[37] Hadikhani, P., Borhani, N., H. Hashemi, S. M., & Psaltis, D. (2019). Learning from droplet flows in microfluidic channels using deep neural networks. *Scientific reports*, *9*(1), 8114.

[38] Erickson, D. (2005). Towards numerical prototyping of labs-on-chip: modeling for integrated microfluidic devices. *Microfluidics and Nanofluidics*, *1*(4), 301-318.

[39] Vedula, E. M., Isenberg, B. C., Santos, J., Lai, W., Lewis, D. J., Sutherland, D., ... & Borenstein, J. T. (2022). Multilayer scaling of a biomimetic microfluidic oxygenator. Asaio Journal, 68(10), 1312-1319.

[40] Liu, J., Lv, Y., Li, X., Feng, S., Yang, W., Zhou, Y., & Tao, S. (2023). Magnetic-optical dual functional Janus particles for the detection of metal ions assisted by machine learning. *Smart Molecules*, *1*(2), e20230006.
[41] Dai, M., Xiao, G., Shao, M., & Zhang, Y. S. (2023). The synergy between deep learning and organs-on-chips for high-throughput drug screening: a review. *Biosensors*, *13*(3), 389.
[42] Muthamilselvan, S., Ramasami Sundhar Baabu, P., & Palaniappan, A. (2023). Microfluidics for Profiling miRNA Biomarker Panels in AI-Assisted Cancer Diagnosis and Prognosis. Technology in Cancer Research & Treatment, 22, 15330338231185284.
[43] Serrano, D. R., Luciano, F. C., Anaya, B. J., Ongoren, B., Kara, A., Molina, G., ... & Lalatsa, A. (2024). Artificial intelligence (AI) applications in drug discovery and drug delivery: Revolutionizing personalized medicine. *Pharmaceutics*, *16*(10), 1328.
[44] Bhaiyya, M., Panigrahi, D., Rewatkar, P., & Haick, H. (2024). Role of Machine Learning Assisted Biosensors in Point-of-Care-Testing For Clinical Decisions. *ACS sensors*, *9*(9), 4495-4519.
[45] Braz, D. C., Neto, M. P., Shimizu, F. M., Sá, A. C., Lima, R. S., Gobbi, A. L., ... & Oliveira Jr, O. N. (2022). Using machine learning and an electronic tongue for discriminating saliva samples from oral cavity cancer patients and healthy individuals. *Talanta*, *243*, 123327.
[46] Williams, D. F., Rahimi, N., Smay, J. E., & Hemmati, S. (2023). Optimizing silver nanowire synthesis: machine learning improves and predicts yield for a polyol, millifluidic flow reactor. *Applied Nanoscience*, *13*(9), 6539-6552.
[47] Lee, R. Y., Wu, Y., Goh, D., Tan, V., Ng, C. W., Lim, J. C. T., ... & Yeong, J. P. S. (2023). Application of artificial intelligence to in vitro tumor modeling and characterization of the tumor microenvironment. *Advanced Healthcare Materials*, *12*(14), 2202457.
[48] Nashruddin, S. N. A. B. M., Salleh, F. H. M., Yunus, R. M., & Zaman, H. B. (2024). Artificial intelligence–powered electrochemical sensor: Recent advances, challenges, and prospects. *Heliyon*.
[49] Gunawardena, H., & Silva, N. (2022). Integrated Sensor and Microfluidic System for On-Farm Drug Screening in Precision Livestock Farming. *Journal of Sustainable Urban Futures*, *12*(12), 30-37.
[50] Hua, M. Z., Liu, J., Roopesh, M. S., & Lu, X. (2024). Microfluidic Optical Aptasensor for Small Molecules Based on Analyte-Tuned Growth of Gold Nanoseeds and Machine Learning-Enhanced Spectrum Analysis: Rapid Detection of Mycotoxins. *ACS sensors*.
[51] Tavakoli, H., Zhou, W., Ma, L., Guo, Q., & Li, X. (2020). Paper and paper hybrid microfluidic devices for point-of-care detection of infectious diseases. *Nanotechnology and Microfluidics*, 177-209.
[52] Walgama, C., Nguyen, M. P., Boatner, L. M., Richards, I., & Crooks, R. M. (2020). Hybrid paper and 3D-printed microfluidic device for electrochemical detection of Ag nanoparticle labels. *Lab on a Chip*, *20*(9), 1648-1657.
[53] Duong, L. H., & Chen, P. C. (2019). Simple and low-cost production of hybrid 3D-printed microfluidic devices. *Biomicrofluidics*, *13*(2).
[54] Vo, T. N. A., & Chen, P. C. (2022). Maximizing interfacial bonding strength between PDMS and PMMA substrates for manufacturing hybrid microfluidic devices withstanding extremely high flow rate and high operation pressure. *Sensors and Actuators A: Physical*, *334*, 113330.
[55] Kojić, S., Birgermajer, S., Radonić, V., Podunavac, I., Jevremov, J., Petrović, B., ... & Stojanović, G. M. (2020). Optimization of hybrid microfluidic chip fabrication methods for biomedical application. *Microfluidics and Nanofluidics*, *24*(9), 66.
[56] Wang, L., Li, B., Wang, J., Qi, J., Li, J., Ma, J., & Chen, L. (2022). A rotary multi-positioned cloth/paper hybrid microfluidic device for simultaneous fluorescence sensing of mercury and lead ions by using ion imprinted technologies. *Journal of Hazardous Materials*, *428*, 128165.
[57] Yan, S., Zhang, J., Yuan, D., & Li, W. (2017). Hybrid microfluidics combined with active and passive approaches for continuous cell separation. *Electrophoresis*, *38*(2), 238-249.
[58] Kim, U., Oh, B., Ahn, J., Lee, S., & Cho, Y. (2022). Inertia–acoustophoresis hybrid microfluidic device for rapid and efficient cell separation. Sensors, 22(13), 4709.
[59] Xiang, J., Qi, J., Hu, D., Wang, C., Wang, L., Wu, Y., ... & Chen, L. (2024). Molecularly imprinted metal-organic frameworks assisted cloth and paper hybrid microfluidic devices for visual detection of gonyautoxin. *Journal of Hazardous Materials*, *469*, 133969.
[60] Tran, Tai Duc, et al. "DNA-copper hybrid nanoflowers as efficient laccase mimics for colorimetric detection of phenolic compounds in paper microfluidic devices." *Biosensors and Bioelectronics* 182 (2021): 113187.