MASTER THESIS PRESENTATION

Hybrid Machine Learning Characterization and Parameter Space Analysis using Interactive Visualization for Analyzing the Quality of Virtual Nonwovens

11th February, 2021

Viny Saajan Victor
Department of Computer Science

Supervisors: **Prof. Dr. Heike Leitte, Dr. Andre Schmeißer**
Technische Universität Kaiserslautern, Germany
Fraunhofer ITWM, Kaiserslautern, Germany

TECHNISCHE UNIVERSITÄT KAISERSLAUTERN

visual information analysis

Simulating Production Process and Virtual Production of Nonwovens

Fiber Dynamics Simulation Tool (Fraunhofer ITWM):

- simulates fibers in turbulent flows.

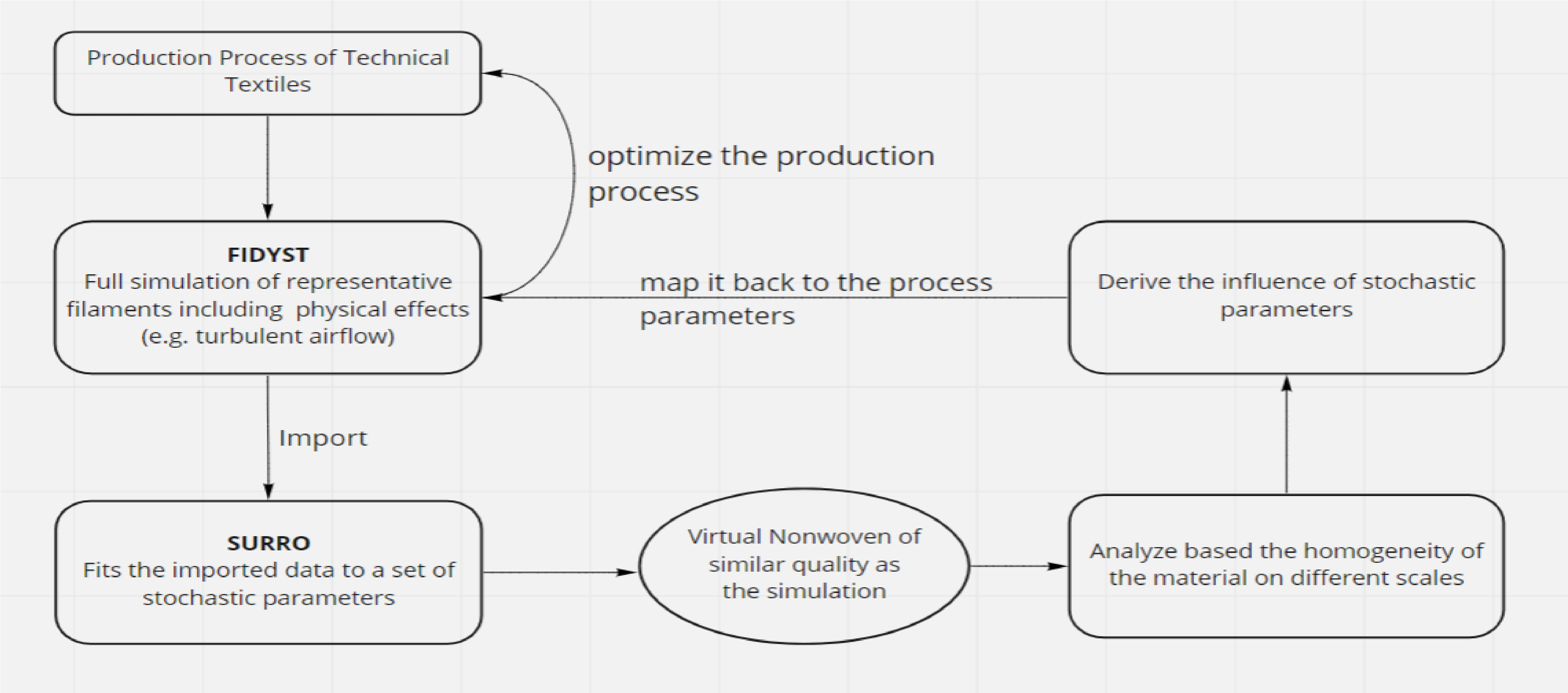
Software Surro (Fraunhofer ITWM):

- generates large-scale virtual nonwoven structures.
- mathematical defined by stochastic differential equation.

Purpose:

- used to optimize the geometry of production plant and operating conditions.
- improve product quality, reduced energy and raw material consumption.

Analysis of Nonwoven production processes using FIDYST and SURRO



Motivation

- **Reduce the computation time and resources:**

  - virtual production of large nonwovens consumes more time and resources.

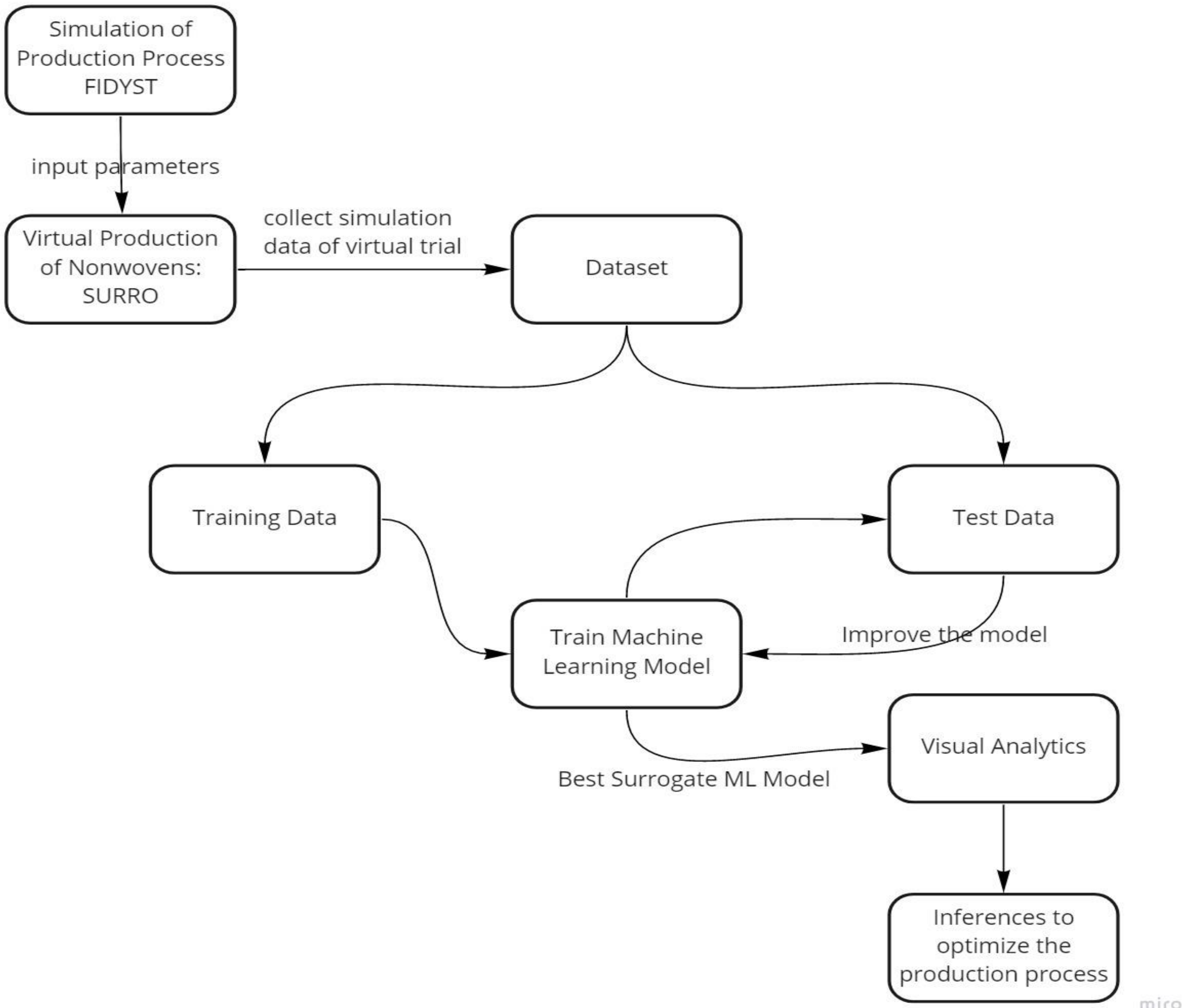
  - require a model that maps the stochastic parameters to the product quality in real time.

- **Visual assistance in analyzing the quality of the nonwovens:**

  - require visual interface which helps the use in parameter space analysis for finding the desired quality of the end product.

  - helps the user to understand the influence of process parameters on the product quality by analyzing in influence of the stochastic parameters.

Thesis Structure



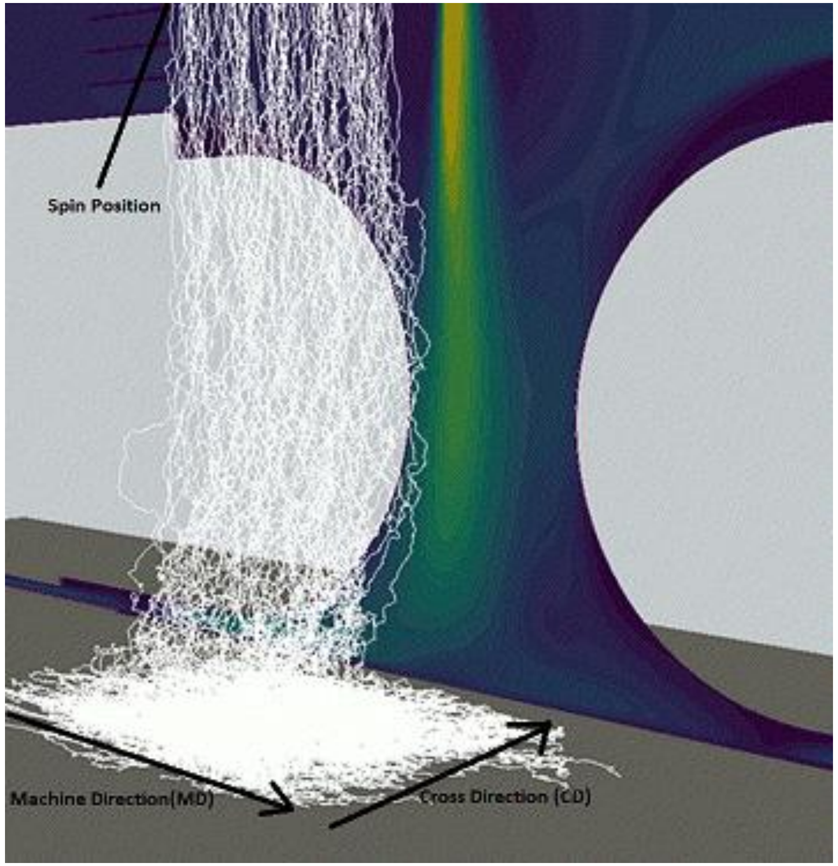
Input and output features for Machine Learning Models

Input Features:

- Sigma_1 **:** standard deviation of normal distribution of fiber position in machine direction without the belt movement, range: [1-50].
- Sigma_2 **:** standard deviation of normal distribution of fiber position in cross direction without the belt movement, range: [1-50].
- A **:** random effects of the production process (eg. Influence of turbulant flow , fiber-fiber contacts etc), range: [1-50].
- BeltSpinRatio **:** ratio of spinning speed and belt speed, range: [0.01-0.25].
- SpinPositionsPerMeter **:** number of spin positions per meter, range: [200-10000].

Output Features:

Coefficient of Variation(CV): Std/Mean : homogeneity : at 7 resolutions.



Nonwoven Sample Simulation Setup

Sample Region Size Selection

Purpose:

- reduce the computation time and memory.
- only simulate the fibers that overlap with sample region.

Selected region sizes:

5cm*5cm, 15cm*50cm, 25cm*50cm : Evaluated with the initial database with 3125 rows.

Inferences:

- Larger the region size -> Lesser the statistical uncertainty.
- Sampling the same parameter setting multiple times reduces the uncertainty.

Choice Made: Sample region size : 25cm*50cm, Number of runs: 5

Nonwoven Sample Simulation Setup

Construction of Nonwoven samples based on Sample Size

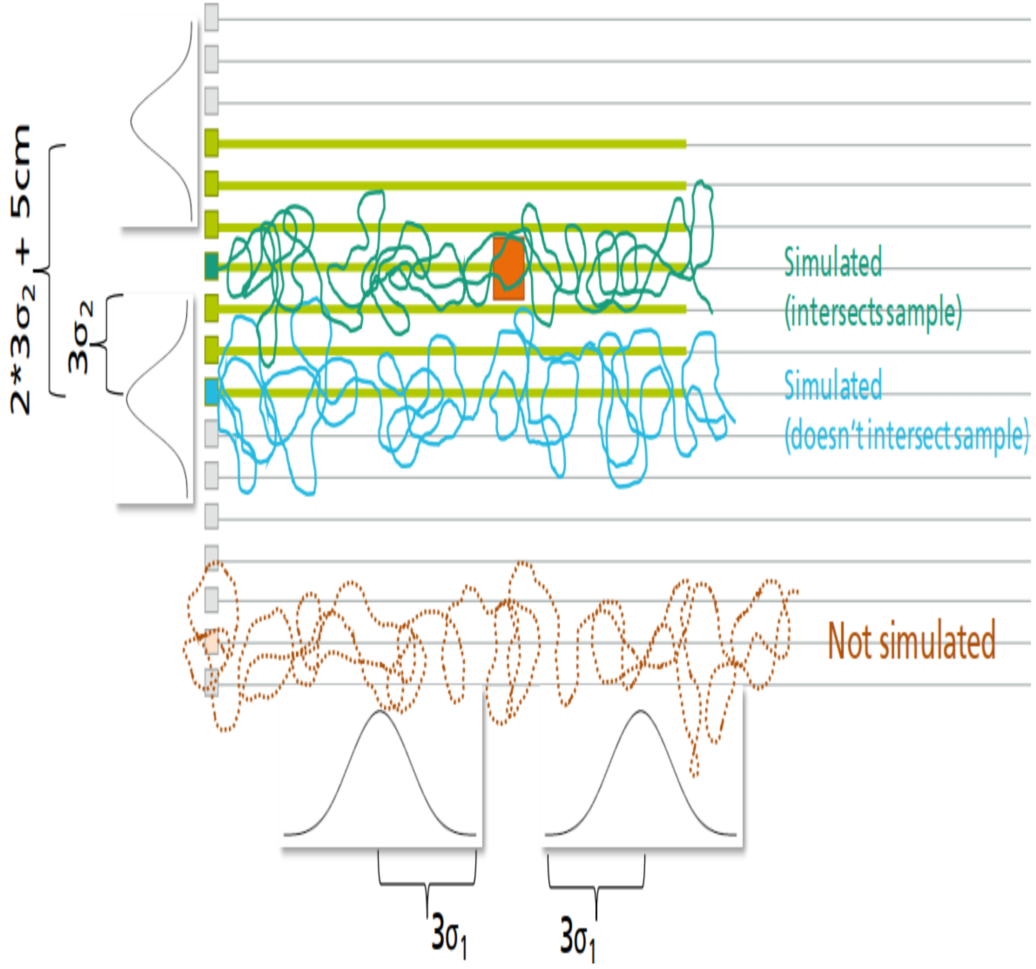
Values lies with 2, 2.5 and 3 std away from

mean in normal distribution are 95%, 99% and 99.7%.

Hence we simulate:

[2.5σ2−3σ2] of nonwoven in cross direction around sample region.

Bit more than 3σ1 in machine direction around sample region



Dataset Creation

Input Database

- smaller number of inputs on a lattice, add additional randomly distributed inputs.
- 50,000 Latin Hypercube Samples(LHS) : distributes samples evenly over sample space.
- 12348 combination of discrete samples.
- total number of rows = (50,000+12348)*5 = 311740

	Sigma_1	Sigma_2	A	BeltSpinRatio	SpinPositionsPerMeterInverse	RandomSeeds
0	1.000000	1.000000	1.000000	0.010000	200.000000	[1213095470, 442944496, 634841805]
1	1.000000	1.000000	1.000000	0.010000	200.000000	[1642774584, 414207576, 1611613430]
2	1.000000	1.000000	1.000000	0.010000	200.000000	[1248361643, 518153081, 1721992811]
3	1.000000	1.000000	1.000000	0.010000	200.000000	[1834129115, 952397124, 1184676440]
4	1.000000	1.000000	1.000000	0.010000	200.000000	[1274595782, 707387979, 1557540557]
...	...	...	...	...	...	...
311735	3.513361	20.232137	23.026302	0.180353	837.352405	[1003094736, 591635343, 1961093723]
311736	3.513361	20.232137	23.026302	0.180353	837.352405	[953213396, 1283929761, 639338037]
311737	3.513361	20.232137	23.026302	0.180353	837.352405	[538647705, 364740067, 1005850105]
311738	3.513361	20.232137	23.026302	0.180353	837.352405	[1442998392, 648729299, 2074783063]
311739	3.513361	20.232137	23.026302	0.180353	837.352405	[635667682, 806371089, 1143299561]

311740 rows × 6 columns

Output Database:

- Input rows + CV values at 7 resolutions.
- Divided the input data into 16 batches of 20,000 rows and ran the simulations.

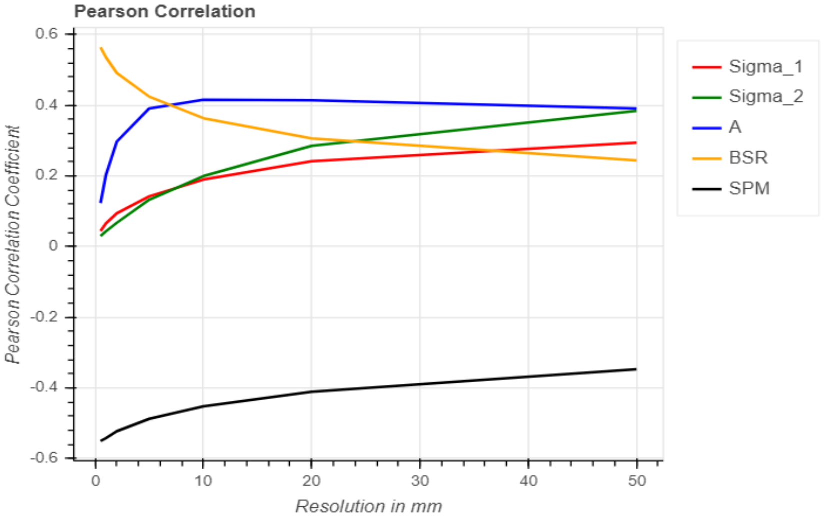
Exploratory Data Analysis and Data Preperation

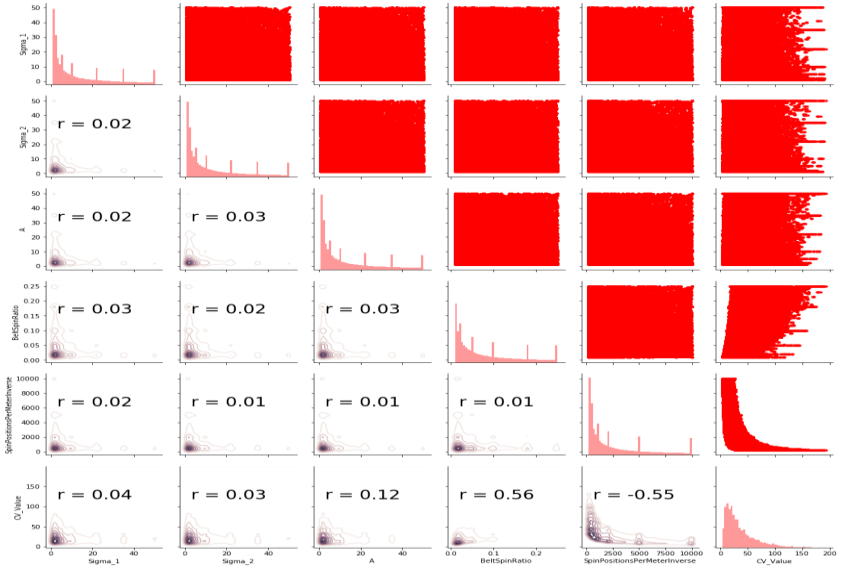
EDA

- Pearson correlation between the input and output features.
- Pair plots for visualizing.
- Sigma_1, Sigma_2, A and BeltSpinRatio positively correlated and SpinPositionsPerMeter is negatively correlated to the output features.

Data Preparation

- data cleaning : removal of random seeds.
- feature scaling: measurements of different units : standardizing the data(rescale with mean=0, and std=1)
- data splitting : training, validation and testing.





Regression Models

Problem setting

5 continuous/discrete input features used to predict 7 continuous output features. Multiple input, multiple output regression.

Model selection criteria

accuracy,interpretability, scalability, confidence on predictions and application domain.

Metrics used for evaluation

Mean Absolute Percentage Error(MAPE).

Mean Squared Error (MSE).

Co-efficient of Determination (R^2 Score).

$$MAPE = \left(\frac{1}{n} * \sum_{i=1}^{n} \frac{y_i - \hat{y}_i}{y_i} \right) * 100$$

$$MSE = \frac{1}{n} * \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$R^2(y, \hat{y}) = 1 - \left(\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y}_i)^2} \right)$$

Here, n refers to the number of data points, y_i and $\hat{y}_i$ refer to the actual and predicted values respectively for the data point i and $\bar{y}$ is the mean value of the data points.

Regression Models

Regression Model	MAPE(%)	MSE	R2-Score
Multi-Linear Regression	94.57	94.55	0.60
Polynomial Regression (Degree = 9)	8.62	1.79	0.99
Random Forest Regression	9.84	1.81	0.98
Bayesian Polynomial Regression (Degree = 9)	7.72	1.51	0.98
Neural Network Regression	3.76	0.35	0.99

Regression Models

Neural Network Regression

- network of functions to understand and translate a data input to desired output .
- Hyper-parameter optimization for best Neural Network Architecture
  * Number of Hidden Layers = [1-5]
  * Number of nodes in each layer = [8-1024 with increment of 8]
  * Activation functions = [relu, sigmoid, tanh]

Resolution (in mm)	MAPE(%)	MSE	R2-Score
0.5	0.80	0.50	1.00
1	1.02	0.47	1.00
2	1.38	0.39	1.00
5	2.18	0.27	1.00
10	3.33	0.27	1.00
20	5.31	0.25	0.99
50	12.31	0.27	0.97
All	3.76	0.35	0.99

Regression Models

Neural Network Regression

- selected network architecture based on accuracy.
- selected activation function: relu

Automated ML Analysis

- plug-and-play model for future integrations of novel unseen data, regression models, new evaluation metric etc.

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
dense (Dense)                (None, 256)               1536
_________________________________________________________________
dense_1 (Dense)              (None, 512)               131584
_________________________________________________________________
dense_2 (Dense)              (None, 512)               262656
_________________________________________________________________
dense_3 (Dense)              (None, 256)               131328
_________________________________________________________________
dense_4 (Dense)              (None, 768)               197376
_________________________________________________________________
dense_5 (Dense)              (None, 7)                 5383
=================================================================
Total params: 729,863
Trainable params: 729,863
Non-trainable params: 0
```

ML_Model_Metrics.csv

Row: 4 Column: 34 Characters: 150

```
1  Regressor,MSE,R2_Score,MAPE(%)
2  Random Forest Regressor,1.81,0.98,9.84
3  Neural Network Regressor,0.35,0.99,3.76
4  Bayesian Regressor,1.51,0.98,7.72
```

Visual Analytic Tool

Main Objectives

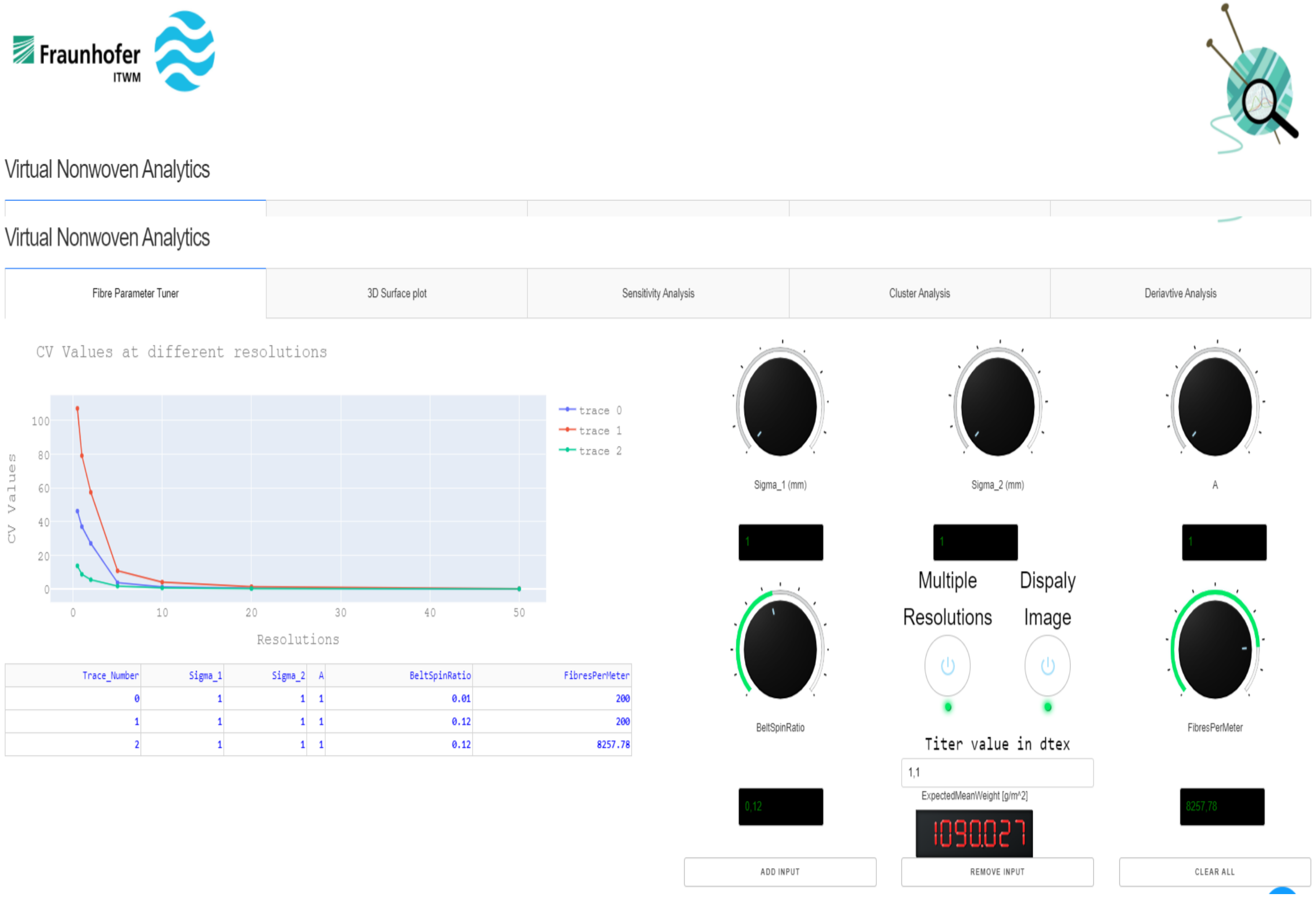
- understanding the relation between input and output parameters and the sensitivity of the output values to minor changes in the input parameters.
- visually aid the users to reach their optimal set of input parameters which result in the desired output.
- show patterns, clusters and anomalies in the data for better understanding of the data.
- data navigation in real-time.

Visual Analytic Tool

Parameter Tuner

- analysis of the output for all the res.
- comparison between more than one parameter setting (add, delete, clear).
- analysis of the output for individual resolutions over a range.
- expected mean weight calculation.

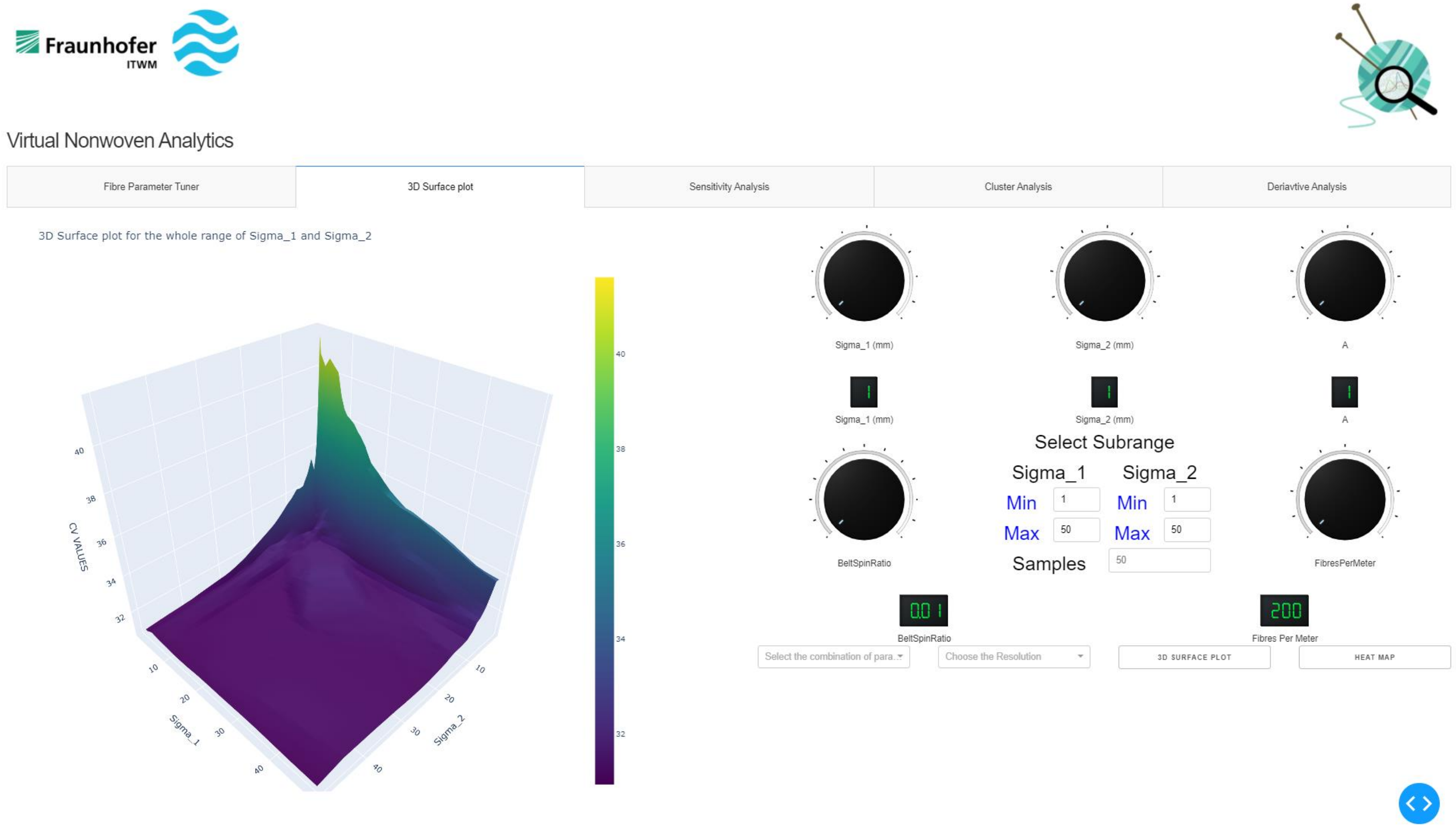
  Expected Mean Weight = titer * SPM* BSR [titer->weight per unit legth]



Visual Analytic Tool

3D Surface Plot

- understanding the influence of input parameters by freezing combination of parameters.
- Dynamic update of the graph with 2500 values in real-time.
- Sub-range selection for "Details on Demand" Analysis.
- Analysis for individual resolutions.

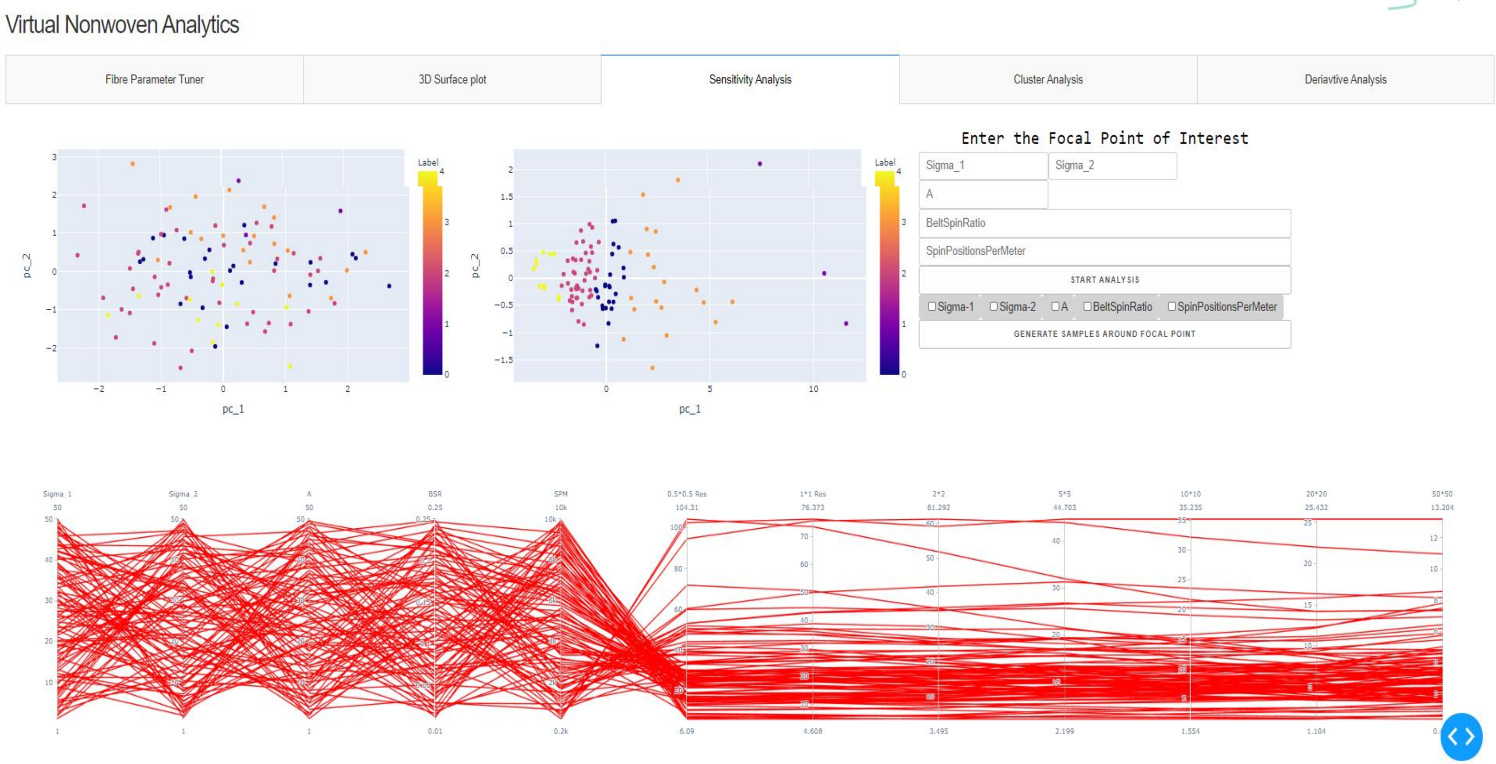


Visual Analytic Tool

Sensitivity Analysis

- to determine the robustness of the output values against small changes in the input parameter settings.

- components : input space graph, output space graph and parallel plot.

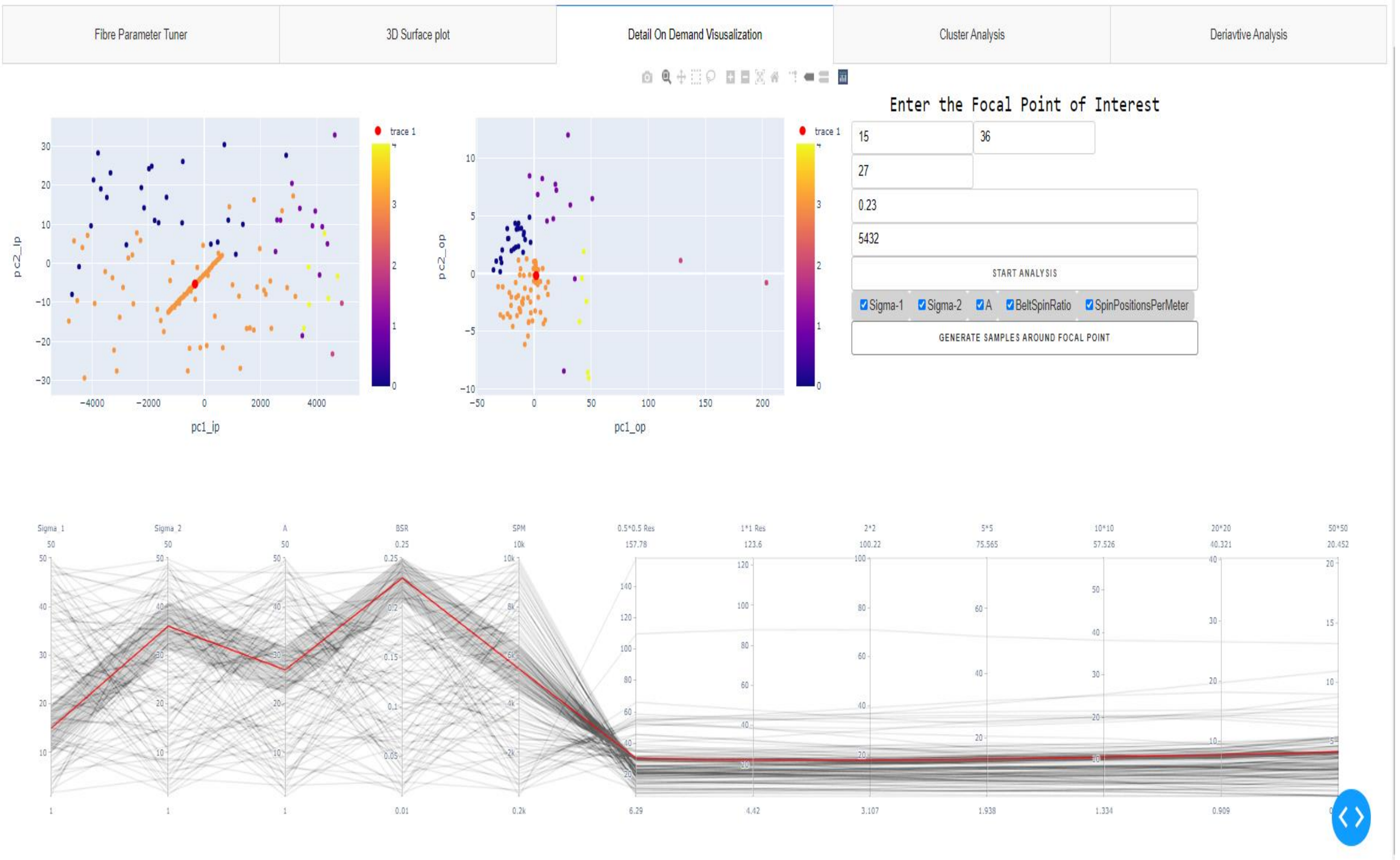


Visual Analytic Tool

Sensitivity Analysis

The Analysis:

- Selection of Focal Point
- Selection of area and population of local neighbourhood.
- Updating the local neighbourhood (local and global update).
- Apply changes on the components



Visual Analytic Tool

Cluster Analysis:

- Partition of output dpace (PCA + GMM clustering).
- Assign a Quality Measure to the Partitioned Data (Average CV value).
- Mapping the partitions to the input space (Box-plot for analysing The distribution of input features).

Visual Analytic Tool

Partial Derivative Analysis:

- partial derivatives of the output with respect to the chosen input at fixed points.
- central finite difference as an approximation for the partial derivatives.
- range slider for the user to select the region of interest.
- Analysis for individual resolutions.

Inferences

- Optimal vs Non-sensitive Parameter Setting using sensitivity analysis.
- Desired resolutions for differentiating nonwovens:
  * higher resolutions do a better job in classifying nonwoven materials as "good" or "bad"compared to lower resolutions.
- Local minima in optimizing the process parameters.
- Outliers in Clusters.

Future Scope

- Point out the exact resolution that differentiate the behaviour of the materials.
- Other metrics to differntiate between the clusters.
- Incorporate the confidence of the machine learning model in the tool using theBayesian regression model.
- utilize the SURRO simulation images instead of CV values and train these images over a Convolutional Neural Networks to find a measure for determining the quality of the material.

THANK YOU