1 Introduction

Machining is one of the most widely used manufacturing processes. The cost of machining contributes to greater than 15% of the total value of the manufactured part [1]. Enhancement of the machining performance by controlling machining parameters has been a long-cherished goal of researchers. There have been numerous attempts in modelling and optimisation of different machining processes. As a result, a lot of data has already been generated and the size of data is continuously increasing. The challenge is to effectively utilise this data for enhancing the overall productivity of an industry. Recently, a renewed interest has been shown all over the world in data analytics; particularly, big data has become a buzzword [2, 3].

Prediction of cutting force, surface roughness and tool wear in machining keep attracting the attention of researchers. Surface roughness has a long bearing on the quality of a product. Cutting forces influence dimensional accuracy, surface integrity, tool life and energy consumption. Tool wear modelling is very important for replacing the tool before it breaks or produces an inferior quality product. A number of analytical and numerical methods have been developed to predict the surface roughness and cutting forces as a function of process parameters [4, 5]. Attempts have also been made to predict tool wear during machining based on online signals from sensors [6, 7]. However, these methods could attain only limited success and prompted the researchers to develop soft computing-based methods [8, 9]. Soft computing-based methods such as neural networks require a large amount of experimental data. With machining industries working in isolation and devoid of a well-established mechanism for data storage, data scarcity has been a significant hindrance to the effective utilisation of soft computing based methods. Fortunately, thanks to progress in information technology, industries are headed towards implementing artificial intelligence (AI) in their activities. Industries are trying to get smart and intelligent by incorporating the technological concepts of cloud computing, cyber-physical system (CPS) and Internet of Things (IoT). These have been identified as the backbone of Industry 4.0 and Made in China 2025 [10]. It appears that data scarcity will no longer be a problem and proper analytics of big data can be a viable substitute for analytical and numerical methods.

The objective of this paper is to sensitise the researchers in the application of big data and analytics in machining. The literature already contains several articles suggesting the prediction of performance parameters in machining using various techniques such as regression and neural network models. However, the biggest challenge is to acquire reliable data of different varieties in sufficient quantity. This paper mainly focuses on this issue, i.e. how a reliable database repository can be prepared that will enable the modelling and optimisation of machining processes. In order to present a proper framework for this task, an example of main cutting force prediction in turning is taken, while fully acknowledging that in a real scenario, the captured data will be supposed to carry out a prediction of a number of parameters. Also, the data will be quite voluminous, unlike those used in this paper to illustrate the proposal.

Before presenting the strategy for the capturing of useful data and prediction based on analytics in Section 3, challenges in the modelling and optimisation of machining followed by a brief review on data-driven approaches in machining are presented in Section 2. Four typical case studies are presented in Section 4. Section 5 concludes the paper.

2 Data-driven approaches in machining: a brief review

Machining has always been one of the most important processes of manufacturing. Despite the researchers taking a keen interest in machining, it is one of the least understood processes [11]. Even today, attempts are being made to develop analytical models for machining [12]. However, due to several simplifying assumptions, the prediction capability of analytical models is limited. Most of the time, constant material properties are taken; significant effect of temperature and strain-rate is not accounted for. Friction is also incorporated in the model in a highly simplified manner, usually through Coulomb’s model. However, it is well-established that Coulomb’s model is inappropriate in the processes involving plastic deformation. Three-dimensional state of stress is also not considered in most of the analytical models, as it is next to impossible to solve all the three-dimensional governing differential equations analytically. The finite element method (FEM), a numerical technique for solving differential equations, provides...
some hope. However, the FEM model can be successful only if the phenomena involved in machining are properly understood. Astakhov [13] analysed the reasons for the poor performance of machining process models. In particular, he refuted the claim of high strain rates [14, 15].

A semi-analytical method for the prediction of cutting forces is mechanistic modelling. Cutting force estimation using mechanistic modelling involves cutting edge discretisation, evaluation of the force components at each discrete element based on cutting constants and edge coefficients, and summing up the obtained force components at each discrete element to evaluate the resultant cutting force. Hence, mechanistic models provide somewhat accurate results and are still in vogue. For example, recently, it has been applied to predict cutting force in the orthogonal machining of an aluminium alloy with cutting speed, feed and edge radii as the input parameters [16]. Nevertheless, it may be a good idea to combine mechanistic and data-driven models. In the combined models, the prediction can be made with mechanistic models but the coefficients of the model can be updated with the feedback of data. In fact, since long, researchers have been stressing on the need to develop an exhaustive database in the machine and associated challenges in procuring data [17].

As against physics-based modelling such as mechanistic and FEM modelling, data-based modelling uses soft computing tools for performance prediction and optimisation [18, 19]. A detailed review of the performance prediction in machining using soft computing has been carried out by Chandrasekaran et al. [20]. One of the pioneering works in the field of performance prediction in machining using soft computing was carried out by Rangwala and Domfeld [21]. They used the feed-forward neural network approach in the turning process, where cutting speed, feed rate and depth of cut were the input variables and cutting force, power consumption, temperature and surface finish were the output variables. Azouzi and Guillot [22] researched predicting the surface finish and dimensional deviation in turning using a neural network. The neural network was trained based on the best combination of the input parameters obtained using the proposed sensor selection and fusion model. Feed, depth of cut, radial force and feed force were the input parameters for the prediction of surface finish and dimensional deviation. Rubbad et al. [23] used neural networks for predicting the surface roughness and dimensional deviation in a turning process. Kohli and Dixit [24] suggested a systematic procedure for the training of a multi-layer perceptron neural network that could be effective even when a limited amount of data was available. They also suggested a methodology to obtain the lower and upper bound estimation of surface roughness apart from the most likely estimate. Due to relatively faster training, radial basis function (RBF), neural networks have also been used in machining [25, 26] based on several empirical equations extracted from machining data books and literature. Later on, researchers observed that employing real-time optimisation strategy with continuous learning is required, as the performance of machining processes is time-dependent [31]. Chandrasekaran et al. [31] suggested to carry out machining optimisation with the inclusion of an AI module that was continuously updated with new information.

Wang and Jawahir [32] were one of the early researchers who used the concept of web-based manufacturing in the area of metal cutting. They developed an interactive system in optimising the cutting conditions by the genetic algorithm in milling operation. Later on, Zheng et al. [33] attempted to achieve agile manufacturing in turning by developing a web-based interface for process parameter selection in order to reduce the overall life cycle cost. In a web-based environment, the cloud can prove to be a platform for providing the computing services and warehouse of data, as suggested by Chandrasekaran et al. [34]. In web-based manufacturing, a need arises for controlling the data having a high volume, high velocity and having high variety [35].

With the emergence of the fourth industrial revolution, there has been an increased tendency of incorporating sensor-based technologies for better monitoring and control. This has resulted in an unprecedented increase in data that has been termed as Industrial Big Data [36]. There is an urgent need to quickly process the accumulated data with the help of data analytics. It would aid in the decision making process with enhanced monitoring, control and optimisation of the industrial processes [37]. Belhadi et al. [38] carried out a detailed review of the use of big data analytics in manufacturing processes and presented some case studies implementing the big data.

Although researchers have started using the data analytics approaches in manufacturing, e.g., in the quality prediction of steel [39] and condition monitoring based maintenance of smart systems [40], there are few papers on the use of big data analytics in the area of machining. Some noteworthy works include the development of data-driven neural network model for the prediction of power consumption in turning based on a set of input parameters provided [41], data-driven energy-efficient computer aided process planning [42], data-driven machining optimisation [43] and modelling for cutting force estimation during turning operation using data-driven finite element method [44].

Lenz et al. [45] felt that the implementation of data analytics in a manufacturing organisation should be cross-departmental rather than departmentalised analytics. It is required for an increased level of transparency, better data quality, reduced wastage of resources, etc. Researchers are now more interested in moving towards collaborative data solutions [46]. Small and medium scale industries would be tremendously benefited if there is any provision of sharing of data. Analytics on a common cloud-based platform improves the decision-making process resulting in reduced monetary loss for all the stakeholders. Still, there is no work that addresses the issues of outliers and dynamically updating the database for the prediction of important machining parameters such as cutting force or surface roughness.

Chandrasekaran et al. [20] have identified six research issues in soft computing-based modelling. Two of them are eliminating outliers in the data and predicting the performance in the form of a confidence interval. The present work focuses on these two issues as a part of the framework for data-based modelling.

### 3 Proposed framework for data collection and machining force prediction in turning

Based on the review of the literature presented in the last section, the following issues have been identified in the context of a proper methodology for data collection in order to create a reliable data bank:

- **Developing a proper methodology for data collection and building a reliable data bank for manufacturing industries is the need of the hour.**
- **Care has to be taken to avoid storing unnecessary machining data in the database in order to avoid data explosion.**
- **Data should be filtered to understand the effect of vibration, tool wear and sensor fault.**

Realising the issues in data-driven approaches in machining, this paper addresses them in the following ways:
• This study takes care of the long-awaited need of the machining industries to acquire a reliable methodology for data collection that will ultimately create a data bank. The developed data bank can be used by machining industries across the globe. Initially, the approach would be to utilise the past machining data present in the literature to develop a data bank called Central Database Repository (CDR) here. In due course of time, with the participation of machining industries, the data will increase in the CDR and eventually become Big Data.

• In this study, a new concept of dynamic reliability is introduced, which is related to the veracity of data. Dynamic reliability gets updated with each new information. It helps in building a highly reliable data bank, as discussed in Section 3.1.

• In order to take care of the uncertainties associated with machining operations, a new concept of mini-repository is introduced. Mini-repository keeps all doubtful data. For an illustration of this concept, multiple-linear regressions are adopted for the performance prediction (cutting force in turning) in this study. (Neural network or any other soft computing method can also be used for performance prediction.) The amount of data in the mini-repository and the deviation from the predicted result jointly determine whether to transfer the data to CDR or not. The detailed description is provided in Section 3.2.

• To eliminate duplicate or nearly similar data in the CDR, a concept based on Cook’s distance is being used. Cook’s distance ensures to keep only the influential or necessary data points. Cook’s distance is described in detail in Section 3.2.

3.1 Overall plan of the proposed framework

It has been observed that a lot of machining data is present in the literature, but the same is not being effectively utilised. If the data from various sources are collected and stored in one place, it can be used for the prediction of machining performance. This work attempts to develop a database repository that would impart extracted knowledge at the shop floor or the process planning section and provide suitable instructions.

A conceptual plan for sharing the data among factories is depicted in Fig. 1. The proposed CDR placed in a cloud server provides support both for data storage as well as for computing facilities. The computing facility is achieved in its Infrastructure as a Service (IAAS) mode. The main focus of this work is towards data acquisition, drawing inferences from the data along with a new concept of dynamic reliability attached to inference and updating the database avoiding data-explosion. The value of the data and the associated reliability are the two major aspects of data that are projected in this article. The CDR initially contains data obtained from various reliable sources like literature, including machining handbooks; with time, the volume of data and its (dynamic) reliability keep on changing.

A scheme for data collection and prediction is shown in Fig. 2. The user enters the process parameters in a computer, which searches the CDR and uses suitable codes to estimate the machining force. Estimation of machining force can be based on regression, neural networks, mechanistic model or any such method or the combination of methods. Section 3.4 describes the application of regression models for estimation of most likely, upper and lower estimates of the main cutting force. If the prediction model is not available for the desired tool-work combination, a rough estimate can still be made with the model for a similar tool-work combination. A suitable correction factor can be employed for reducing the model error, based on the shop floor feedback. Any new experimental data-entry is included in the CDR, assuming a 50% chance of it being accurate. Thus, the reliability of 0.5 is attached to a new data-entry. The assumption of attaching the reliability of 0.5 to a new data-entry is based on simple probability theory. Whenever any new information arrives, there can be two possibilities: (1) data-entry truly represents the process or (2) data-entry is erroneous. Hence, there must be some basis for either accepting or rejecting a new data-entry.

Considering the equal possibility of two events:

\[ P(\text{Data is accurate}) = \frac{1}{2}; \quad P(\text{Data is inaccurate}) = \frac{1}{2}. \]  

Therefore, a new data-entry is stored by assigning a reliability of 0.5. In other words, there is a probability of 0.5 of it being correct. Nevertheless, a value other than 0.5 may be assigned if there is any prior information based on the confidence of experts.

Apart from the information about cutting-tool and work-piece material, a variety of information can be collected. Often the information is incomplete. Several articles in the literature predict the cutting force based on only three parameters – cutting speed,
feed and depth of cut. However, at times data about vibration and tool wear are also available. Besides, modern adaptive control computer numerical control (CNC) machines can have data from several sensors, e.g. acoustic emission, temperature and spindle motor current. In order to avoid complicacy in the description, this paper focuses on the cases when the domain of input variables comprises only cutting speed, feed and depth of cut. Of course, a brief description is provided for handling and utilising other types of time-dependent signals. A detailed flowchart for enhancing the machining performance using Big Data Strategy is shown in Fig. 3 and the corresponding related actions are listed in Table 1. The flowchart provides an outline of the proposed strategy in various uncertain situations that are explained in detail. This framework can be extended to machining optimisation; however, it is not the focus of this article.

3.2 Methodology for data collection when feed, depth of cut and cutting speed are available along with main cutting force

Assume that the following information is available about the machining process: tool material, work-piece material, presence or absence of lubrication, cutting speed, feed, depth of cut and cutting force. There can be different scenarios, which have to be tackled in different ways. Sometimes the information can be entirely new and at other times, similar information may be available. In the sequel, different cases are discussed.

Case 1: Existing cutting tool and work-piece combination. Whenever machining is performed for the existing cutting tool and work-piece combination, i.e. the tool-work combination is already present in the CDR. The process parameters in which machining is taking place are entered and the prediction is made. After the actual machining, the feedback from the shop floor can be collected for updating the prediction model. The following are the two possible scenarios:

Case 1A: The process parameters data laying outside the range of the data in CDR. Whenever the incoming process parameters for the existing cutting tool and work-piece combination are outside the range of data in CDR, the extrapolated prediction is carried out. There is a high probability of the predicted value being inaccurate. The percentage of prediction error depends on the proximity of the predicted process parameter to the range of the data already in the database. Hence, the prediction of extrapolated data should always be made with a caveat. After the completion of the machining operation, shop floor feedback can be obtained. The new information is then stored in CDR with a reliability of 0.5.

Case 1B: The process parameters data laying within the range of the data in CDR. Whenever the incoming process parameters for the existing tool-work combination are within the range of the existing process parameters in CDR, estimation of the cutting force is carried out within 95% confidence interval or any such interval. In case the feedback of the cutting force from the shop floor is within the estimated range, the decision whether to include the incoming data in the CDR or not is taken based on the Cook’s and the Euclidean distances. It is explained in the following paragraphs.

Concept of Cook’s distance: A significant way to measure the influence of observation is by using Cook’s distance. Cook’s distance considers the residual of the output and the leverage of the point while estimating the level of influence. In other words, Cook’s distance combines the value of an outlier and leverage for estimating the level of influence. In this study, the concept of Cook’s distance is used with multiple-linear regression for illustration only. It can be interpreted as a means to indicate the possible locations where it would be good to have more number of data points that can affect future outcomes.

In case the feedback of the cutting force from the shop floor is within the estimated range, the decision whether to include the incoming data in the CDR or not is taken based on the Cook’s and the Euclidean distances. It is explained in the following paragraphs.

Fig. 2 Proposed scheme for sending feedback to CDR based on the actual result from the shop floor

This is an open access article published by the IET under the Creative Commons Attribution License (http://creativecommons.org/licenses/by/3.0/)
\[
D_i = \left( \frac{y_i - y_{i,p}}{(m+1)s} \right)^2 \frac{h_i}{(1-h_i)}.
\]

where \( y_i \) is the actual obtained output for the \( i \)th observation, \( y_{i,p} \) is the corresponding predicted output, \( m \) is the number of predictors (three viz., cutting speed, feed and depth of cut, in this case), \( s \) is the standard error of the estimate and \( h_i \) is the leverage of the \( i \)th observation. The leverage is calculated as

\[
h_i = \frac{1}{n} + \frac{(\ln V_p - \ln V)^2}{\sum_{i=1}^{n}(\ln V_i - \ln V)^2} + \frac{(\ln f_p - \ln f)^2}{\sum_{i=1}^{n}(\ln f_i - \ln f)^2} + \frac{(\ln d_p - \ln d)^2}{\sum_{i=1}^{n}(\ln d_i - \ln d)^2}.
\]

A data-entry is considered to be influential only when the value of Cook’s distance exceeds one [47]. The influential points i.e., the data points having Cook’s distance more than one, will be included in CDR with a 50% reliability. When Cook’s distance does not exceed one, the nearest data point is searched in the CDR and its reliability is improved. This strategy will help in avoiding data-exploration as it does not store non-influential data.

\[
D_i = \left( \frac{y_i - y_{i,p}}{(m+1)s} \right)^2 \frac{h_i}{(1-h_i)}.
\]

where \( y_i \) is the actual obtained output for the \( i \)th observation, \( y_{i,p} \) is the corresponding predicted output, \( m \) is the number of predictors (three viz., cutting speed, feed and depth of cut, in this case), \( s \) is the standard error of the estimate and \( h_i \) is the leverage of the \( i \)th observation. The leverage is calculated as

\[
h_i = \frac{1}{n} + \frac{(\ln V_p - \ln V)^2}{\sum_{i=1}^{n}(\ln V_i - \ln V)^2} + \frac{(\ln f_p - \ln f)^2}{\sum_{i=1}^{n}(\ln f_i - \ln f)^2} + \frac{(\ln d_p - \ln d)^2}{\sum_{i=1}^{n}(\ln d_i - \ln d)^2}.
\]

A data-entry is considered to be influential only when the value of Cook’s distance exceeds one [47]. The influential points i.e., the data points having Cook’s distance more than one, will be included in CDR with a 50% reliability. When Cook’s distance does not exceed one, the nearest data point is searched in the CDR and its reliability is improved. This strategy will help in avoiding data-exploration as it does not store non-influential data.

\[
D_i = \left( \frac{y_i - y_{i,p}}{(m+1)s} \right)^2 \frac{h_i}{(1-h_i)}.
\]

where \( y_i \) is the actual obtained output for the \( i \)th observation, \( y_{i,p} \) is the corresponding predicted output, \( m \) is the number of predictors (three viz., cutting speed, feed and depth of cut, in this case), \( s \) is the standard error of the estimate and \( h_i \) is the leverage of the \( i \)th observation. The leverage is calculated as

\[
h_i = \frac{1}{n} + \frac{(\ln V_p - \ln V)^2}{\sum_{i=1}^{n}(\ln V_i - \ln V)^2} + \frac{(\ln f_p - \ln f)^2}{\sum_{i=1}^{n}(\ln f_i - \ln f)^2} + \frac{(\ln d_p - \ln d)^2}{\sum_{i=1}^{n}(\ln d_i - \ln d)^2}.
\]

A data-entry is considered to be influential only when the value of Cook’s distance exceeds one [47]. The influential points i.e., the data points having Cook’s distance more than one, will be included in CDR with a 50% reliability. When Cook’s distance does not exceed one, the nearest data point is searched in the CDR and its reliability is improved. This strategy will help in avoiding data-exploration as it does not store non-influential data.

\[
D_i = \left( \frac{y_i - y_{i,p}}{(m+1)s} \right)^2 \frac{h_i}{(1-h_i)}.
\]

where \( y_i \) is the actual obtained output for the \( i \)th observation, \( y_{i,p} \) is the corresponding predicted output, \( m \) is the number of predictors (three viz., cutting speed, feed and depth of cut, in this case), \( s \) is the standard error of the estimate and \( h_i \) is the leverage of the \( i \)th observation. The leverage is calculated as

\[
h_i = \frac{1}{n} + \frac{(\ln V_p - \ln V)^2}{\sum_{i=1}^{n}(\ln V_i - \ln V)^2} + \frac{(\ln f_p - \ln f)^2}{\sum_{i=1}^{n}(\ln f_i - \ln f)^2} + \frac{(\ln d_p - \ln d)^2}{\sum_{i=1}^{n}(\ln d_i - \ln d)^2}.
\]

A data-entry is considered to be influential only when the value of Cook’s distance exceeds one [47]. The influential points i.e., the data points having Cook’s distance more than one, will be included in CDR with a 50% reliability. When Cook’s distance does not exceed one, the nearest data point is searched in the CDR and its reliability is improved. This strategy will help in avoiding data-exploration as it does not store non-influential data.

\[
D_i = \left( \frac{y_i - y_{i,p}}{(m+1)s} \right)^2 \frac{h_i}{(1-h_i)}.
\]
Updated reliability $= 0.5(1 + \text{current reliability})$.  

(6)

As an example, if there is a data-entry already in CDR with a reliability of 0.5 and a similar new data-entry is found, the reliability of the data is updated to 0.75. If the next time another similar data-entry is found, the reliability of the data is updated to 0.875. The reliability value keeps getting updated with more similar data, as shown in Fig. 4. For 7 data, reliability becomes more than 0.99. Thus, the present methodology focuses on building a compact, highly reliable and quality database.

Concept of mini-repository: In case the feedback of the cutting force from the shop floor is outside the predicted range, the scheme will look whether the feedback of the cutting force is below the lower limit or above the upper limit. If the cutting force is below the lower limit, the incoming data-entry will be stored in a separate mini-repository. Data in the mini-repository will be merged with the data in the CDR, when the number of the data in the mini-repository becomes equal to the number of data kept in the CDR for a particular tool-work combination. It is accomplished in the following way:

1. If the number of data in the mini-repository becomes equal to the number of data in CDR, the mean value and standard deviation of the unsigned error of all the data which are below the lower limit for a particular tool-work combination is evaluated. They are denoted by $\mu$ and $\sigma$, respectively. All the data that lie above the $(\mu + 2\sigma)$ are considered as outliers and are excluded. The data that lies below the value of $(\mu + 2\sigma)$ are included and are merged with the main data in the CDR by associating a reliability of 0.5 with the merged data.

If the feedback of the cutting force from the shop floor lies above the estimated upper limit, the condition of the cutting tool as well as vibration data, is verified. One module can be prepared that analyses chatter based on the vibration and other supporting data. If the chatter is detected, the data can be stored in a mini-repository pertaining to machining data with chatter. If there is no chatter, but the tool is old, the data pertaining to tool wear is recorded and information is put in another mini repository. If it is ensured that there is no tool wear or chatter, the data is included in the CDR with a 50% chance of the information being accurate.

Mini-repository contains all the data that does not match with the predicted result, thereby taking care of the uncertainties associated with a machining operation (turning for the present case). Separate analyses can be carried out in mini-repository alone to identify the causes of such deviation from the predicted output. This paper is conceptual in nature, details on analysing wear and chatter signals are omitted.

Example: 

Case 2: New cutting tool and work-piece combination. 

Whenever machining is performed for a new cutting tool and work-piece combination, the estimation of cutting force can be done by utilising all the existing models inside the CDR. The model which has the least error based on the feedback obtained from the shop floor can be chosen for further estimation. After the actual machining is carried out, the model can be updated based on the measurement of cutting force. In order to minimise the model error, a suitable correction factor may be estimated as follows:

\[
\text{Correction factor} = \frac{\text{Feedback of the cutting force}}{\text{Predicted value}} 
\]

(7)

This correction factor is multiplied to an estimated value from the existing similar model. The new information with the incoming data-entry will be included in the CDR for a new cutting tool and work-piece combination with a reliability of 0.5.

3.3 Methodology for data collection when other information is available besides cutting speed, feed and depth of cut

Cutting force is affected by the presence of tool wear and chatter. Researchers have used acoustic emission sensors, accelerometer, dynamometer, etc., for capturing the time-varying signals [48]. The time-varying signals can be stored and used for modelling of the cutting force equation. In this paper, it is assumed that the dependence of cutting force on all the process parameters as linear on a logarithmic scale. For example, in order to estimate the cutting force in turning, the following mathematical model can be fitted:

\[
F = KV^\alpha f^\beta d^\gamma C^\delta, 
\]

(8)

where $F$ is the main cutting force, $V$ is the cutting velocity, $f$ is the feed, $d$ is the depth of cut, $w$ is the height of wear land, $t$ is the time elapsed, $a$ is the acceleration and $C$ is the average ring-down count obtained from an acoustic sensor. The exponents and constant $K$ can be obtained through a multiple-regression procedure. The rest of the procedure is similar to that described in Section 3.2. In order not to lose the focus of this paper, further discussion will pertain to the case when only the information about feed, speed and depth of cut is available along with the main cutting force.

3.4 Estimation of cutting force with a 95% prediction interval

Cutting force mainly depends on cutting speed, feed and depth of cut. Assuming that a sufficient quantity of data is available, in a sufficiently small range of process variables, the following mathematical model can be fitted:

\[
F = KV^\alpha f^\beta d^\gamma, 
\]

(9)

**Table 1** List of actions

| Action number | Task |
|---------------|------|
| 1             | Estimate the cutting force with all the existing models in the database repository and choose the one who is having the least error. Reduce the model error further by updating the estimation with a suitable correction factor and include the information in the prediction module for the new tool-work combination. |
| 2             | Extrapolated prediction is carried out with a caveat and the incoming process parameter along with the output details, are included in the corresponding module of database repository. Assign a reliability of 0.5 to these data. |
| 3             | Extract the output from the prediction model from CDR. |
| 4             | Calculate the Euclidean and Cook’s distances. |
| 5             | Include the incoming process parameter and output details in the corresponding module of the database repository. Assign a reliability of 0.5 to these data. |
| 6             | Improve the reliability of the nearest data (based on the Euclidean distance) in the database repository. |
| 7             | Store the data in a separate module for the analysis of tool wear or chatter. |
| 8             | Save the data in a mini repository. |
| 9             | Merge the data with the data in the CDR pertaining to the same tool-work combination, leaving aside outliers (error above $\mu + 2\sigma$ limit). Fit new regression models. |

![Fig. 4 Improvement of reliability with number of data](image-url)
where $F$ is the main cutting force in $N$, $V$ is the cutting velocity in m/min, $f$ is the feed in mm/rev and $d$ is the depth of cut in mm. The exponents $\alpha$, $\beta$ and $\gamma$ as well as constant $K$ need to be obtained through a multiple-regression procedure. It is very difficult to estimate the form of cutting force equations for a wide range of process parameters. However, for a small range of parameters, the dependence of force on the process parameters can be assumed as linear in the logarithmic scale. To assess the validity of this assumption, the coefficient of determination should be estimated; a high value of the coefficient of determination supports the assumption. A similar form of the cutting force equation has been observed in the literature [49, 50]. Taking the natural logarithm on both sides of (9)

$$\ln F = \ln K + \alpha \ln V + \beta \ln f + \gamma \ln d,$$  \hspace{1cm} (10)

The linear equation obtained in (10) is used for the multiple-linear regression analysis, which amounts to solving the following least square error and the model can be fitted as the data frame in ‘R’. After that, the value of the process parameters, i.e. cutting speed, feed and depth of cut are entered. On entering the details, there are two possibilities: (i) The process parameters are within the range of data in CDR and (ii) The process parameters are outside the range of data in CDR. It is already assumed that the entered data assumed here are within range. Let the cutting velocity, feed, depth of cut for the machined surface be 200 m/min, 0.30 mm/rev and 0.80 mm, respectively. Table 2 shows the range of values of cutting speed, feed and depth of cut that is stored in the database for sintered carbide and S55C high carbon steel.

Table 2 Initial range of values for sintered carbide and S55C high carbon steel
| $V$, m/min | $F$, mm/rev | $d$, mm |
|------------|-------------|---------|
| maximum value | 202.6 | 0.32 | 1.25 |
| minimum value | 86.1 | 0.08 | 0.35 |

4.1 Data collection for existing cutting tool and work-piece combination with parameters laying within the range of data in CDR

It is assumed that Factory A is accessing the CDR and it wants to machine S55C high carbon steel with a sintered carbide tool [51]. It is found that the desired cutting tool and work-piece combination is present in the CDR that is initially having 27 observations. All the data are having a 50% chance of being accurate. Hence, all the data are having the reliability of 0.5. Prediction of cutting force is carried out using ‘R’, an open-source statistical programming language. On executing the multiple-regression analysis, it loads the data file from the CDR in the ‘R’ environment to form a data frame using the ‘read’ library package. The two-dimensional table containing the information about the machining process is referred to as the data frame in ‘R’. After that, the value of the process parameters, i.e. cutting speed, feed and depth of cut are entered. On entering the details, there are two possibilities: (i) The process parameters are within the range of data in CDR and (ii) The process parameters are outside the range of data in CDR. It is already assumed that the entered data assumed here are within range. Let the cutting velocity, feed, depth of cut for the machined surface be 200 m/min, 0.30 mm/rev and 0.80 mm, respectively.

$$\ln F = \ln K + \alpha \ln V + \beta \ln f + \gamma \ln d,$$  \hspace{1cm} (11)

where $i$ is a typical data-entry among the total $n$ data entries available for model fitting.

The solution of (11) provides the optimum values of $K$, $\alpha$, $\beta$ and $\gamma$. This can be carried out for different cutting tool and work-piece combinations. Knowing the optimum values of $K$, $\alpha$, $\beta$ and $\gamma$, (9) can be used to estimate the main cutting force. However, it is always better to predict an estimate of an interval, in which the cutting force may lie with 95% confidence. The reliability of all the data is initially assumed to be the same. The value of the reliability gets updated in due course. The reliability value can be utilised as a weight factor in the least square error and the model can be fitted.

The best fit value predicted using (15) is 711 N. The lower limit and upper limit values of the cutting force predicted using (12) are 646 and 782 N, respectively.

Feedback of cutting force is obtained during machining. Suppose that a value of 700 N is obtained, which is within the predicted interval limits. The next step is to find the level of influence that the entering observation has over the regression model. It is evaluated using Cook's distance. Before evaluating Cook's distance in 'R', the columns containing information about the cutting speed, feed, depth of cut and cutting force are extracted from the data frame using the 'subset' function. The resulting data frame is combined with the data-entry, including the feedback about the cutting force. This leads to the formation of a total of 28 data points from the existing value of 27. On using 'cook's.distance' function in 'R', Cook's distance of the data-entry is calculated. Out of the 28 numeric values obtained for this case, the last value is the Cook's distance for the entering data point. Cook's distance comes out to be 0.00019, which is much less than 1. Hence the entering observation is not influential. It is excluded from entering CDR, instead, the reliability of the nearest data point is improved based on the Euclidean distance.

From the data frame in ‘R’ containing the machining information obtained from CDR, the columns containing the information about cutting speed, feed and depth of cut are extracted using the ‘subset’ function. The extracted information from the
data frame is combined with the data-entry to from a data frame of 28 rows and 3 columns. Euclidean distance is then calculated by using the ‘distance’ function in ‘R’, resulting in a 28 × 28 matrix. The minimum value is found to be for the 6th observation. Therefore, the reliability of the sixth observation is improved to 0.75 using (6) and saved in the CDR. The updated number of the data for machining of S55C high carbon steel with a sintered carbide tool is still 27. However, the reliability of the sixth observation gets updated.

4.2 Data collection for existing cutting tool and work-piece combination with any of the process parameters laying outside the range of data in CDR

In this case, it is assumed that Factory B is accessing CDR and it also wants to machine S55C high carbon steel with a sintered carbide tool. The user enters the cutting velocity, feed, depth of cut for the machined work-piece as 300 m/min, 0.30 mm/rev and 0.80 mm, respectively. It can be seen from Table 2 that the value of the cutting speed entered by the user lies outside the range of data in CDR. This indicates the data-entry is outside the range. Hence, the extrapolated results may be provided with a caveat. Suppose the feedback of the cutting force is 680 N. This new information is included in the CDR with a reliability of 0.5 for later use. The updated number of data for machining of S55C high carbon steel with a sintered carbide tool increases by one.

4.3 Data collection for existing cutting tool and work-piece combination, the process parameters laying within the range of data in CDR with the feedback of the cutting force below the acceptable interval limit

Suppose Factory C is accessing CDR and it also wants to machine S55C high carbon steel with a sintered carbide tool, similar to Factory A and Factory B. On executing the multiple-regression model from the CDR, the user enters cutting speed, feed and depth of cut as 200 m/min, 0.30 mm/rev and 0.80 mm, respectively (as in Section 4.1). Therefore, the process parameters are well within the range of data in CDR, as shown in Table 2. Suppose that this time the feedback obtained about the cutting force is 600 N, which is outside the acceptable limits of 646 and 782 N. As the actual cutting force is outside the acceptable limits, the data-entry is stored in a separate mini-repository. In this case, the number of data in CDR is unaffected. The data that is collected in the mini-repository will be merged with the data in CDR based on the aforesaid condition.

4.4 Data collection for new cutting tool and work-piece combination

Suppose Factory D wants to machine the S45C steel bar with a tungsten carbide tool [52]. It could not find this tool-work in CDR. The existing models present in the library are used for the prediction of the cutting force based on the input process parameters provided during machining. In this example, let the user enters cutting speed, feed and depth of cut as 135 m/min, 0.08 mm/rev and 0.6 mm, respectively. Let the feedback of the cutting force obtained from the shop floor be 263 N. All the models are utilised for finding the cutting force. The results are saved in a data frame. The one having the least error is selected for further estimation. Assuming the minimum absolute error comes out in case when the cutting tool is sintered carbide and work-piece is S55C high carbon steel. The value predicted on substituting the entered process parameter in (15) is 272 N. The feedback of the cutting force obtained. In comparison, machining is 263 N. Hence, an absolute error of 9 N is obtained. In order to reduce the error further, a suitable correction is found:

\[
\text{Correction factor} = \frac{\text{Feedback of the cutting force}}{\text{Predicted value}}
\]

\[
= \frac{263}{272} = 0.967.
\]

The inclusion of the correction factor reduces the mean absolute error in predicting the cutting force. Thus, updated (15) along with the correction factor is used for further prediction of cutting force in the case of tungsten carbide and S45C steel bar. All the new data are collected in the CDR for this cutting-tool and work-piece combination. This new data will, of course, get updated with a sufficient number of trials on the shop floor. The aforesaid procedure for estimating the cutting force using regression methodology is illustrative only. There can be several other methods for the estimation of cutting force. The main objective of this research work is to conceptually demonstrate a procedure of collecting data and building a highly reliable data bank, which is termed as the Central Database Repository (CDR) in this case. The challenge is not only to prepare a reliable database for a manufacturing environment that can be easily accessed via the Internet but also to protect it from hackers and malicious users who can manipulate the data stored in the cloud. Therefore, data confidentiality, integrity and availability become an integral part of the proposed CDR. Service providers for CDR need to ensure that there is at least a tested encryption scheme to protect the data. Preventing unauthorised access and periodic data backup are the other necessities. Researchers are thoroughly carrying out research to improve the level of encryption and to provide enhanced security and privacy to the stored data on cloud-based platforms [53, 54]. However, this is not the focus of present work; implementation of data security and privacy in the context of the proposed framework is left as a future task.

Most importantly, the present work focuses on the usage of Cook’s distance to evaluate the level of influence and for keeping only the important and influential data rather than keeping all the non-influential data. A novel feature of the proposed framework is the usage of dynamic reliability, which keeps on updating with the available information.

5 Conclusion

With the advent of Industry 4.0 and increasing emphasis on using data analytics for distributed manufacturing, it is high time to explore the use of past data in enhancing machining performance. The proposed framework provides systematic guidance for data collection and building a highly reliable and compact data bank for manufacturing industries. CDR supplies on demand data to any factory located within reach of the Internet. There is a provision to properly update the information presented in CDR. The information stored in the data bank can be used for performance prediction that would help in reducing unnecessary downtime. This is illustrated for example of cutting force estimation for turning operations. For ease of illustration, a multiple-linear regression is used for the estimation of the main cutting force within a 95% prediction interval. In practice, a neural network, fuzzy set-based or any other prediction model can also be used.

The purpose of this paper is to project the novel idea of CDR, mini-repository and dynamic reliability. The dynamic reliability can also be used as a weight factor for minimising the weighted least square error in multiple-linear regression. Cook’s distance has also been used as a parameter for measuring the level of influence that can be used for filtering out the non-influential data. In other words, the proposed framework illustrates the application of using data analytics in manufacturing conceptually. Hence, deliberately a small size of data has been used and necessary calculations are done using R, an open-source statistical software. It is envisaged that future machine tools will be part of the cyber-physical system and collection of data by the cloud will not be a difficult task.

This paper proposes the concept required for building a highly reliable databank keeping data explosion in mind for the manufacturing industries. In future, an industrial project is planned for real-life implementation of the proposed methodology. The real implementation will involve estimation of surface roughness, tool wear and all components of forces. The scope can include important machining processes such as milling, drilling and grinding. For better accuracy in the estimation of machining performance, one can explore the use of support vector regression, classification & regression trees, along with neural networks.
within the proposed framework. This work is just an early attempt of applying data analytics in machining.

6 References

[1] Merchant, M.E.: ‘An interpretive look at 20th century research on modeling of machining’, Mach. Sci. Technol., 1998, 2, (2), pp. 157–163.
[2] Lazar, N.: ‘The big picture: crowdsourcing your way to big data’, Chance, 2019, 32, (2), pp. 1–5.
[3] Wu, X., Zhu, X., Wu, G.Q., et al.: ‘Data mining with big data’, IEEE Trans. Knowl. Data Eng., 2014, 26, (1), pp. 97–107.
[4] Wang, H., Dong, H., Ke, Y., et al.: ‘Modeling and analysis of force prediction in milling process of unidirectional fiber reinforced polymer composites’, Int. J. Adv. Manuf. Technol., 2019, 101, (9–12), pp. 3073–3080.
[5] Zhang, X., Yu, T., Wang, W.: ‘Cutting forces modeling for micro flat end milling by considering tool run-out and bottom edge cutting effect’, Proc. Inst. Mech. Eng. B, J. Eng. Manuf., 2019, 233, (2), pp. 470–485.
[6] Kuntoğlu, M., Sağlam, H.: ‘Investigation of progressive tool wear for determining of optimized machining parameters in turning’, Measurement, 2019, 140, pp. 427–436.
[7] Liu, M.K., Tseng, Y.H., Tran, M.Q.: ‘Tool wear monitoring and prediction based on signal based sound’, Int. J. Adv. Manuf. Technol., 2019, 103, (9–12), pp. 3361–3373.
[8] Laghari, R.A., Li, J., Laghari, A.A., et al.: ‘A review on application of soft computing techniques in machining of particle reinforcement metal matrix composites’, Comput. Methods Eng., 2019, pp. 1–15, doi: 10.1007/s11831-019-09340-0.
[9] Cherukuri, H., Perez-Bernabeu, E., Selles, M.A., et al.: ‘A neural network approach for chatter prediction in turning’, Procedia Manuf., 2019, 34, pp. 885–892.
[10] Ji, W., Yin, S., Wang, L.: ‘A big data analytics based machining optimisation approach’, J. Intell. Manuf., 2019, 30, (3), pp. 1483–1495.
[11] Usai, E.: ‘Progress in predictive’ turning’, JSME Int. J. Ser. III, Vib. Control Eng. Ind. Eng., 1988, 31, (2), pp. 363–369.
[12] Utsumi, K., Shichiri, S., Sasahara, H.: ‘Determining the effect of tool posture on cutting force in a turn milling process using an analytical prediction model’, Int. J. Mach. Tools Manuf., 2020, 150, pp. 103511.
[13] Astakhov, V.P.: ‘Authentication of FEM in metal cutting’, in Paulo Davim, J (Eds.): ‘Finite element method for manufacturing processes: models, techniques and applications’, John Wiley & Sons, 2013, pp. 32–36.
[14] Kronenberg, M.: ‘Machining science and application’, vol. 217 (Pergamon Press, London, 1996), p. 219.
[15] Stephens, J.: ‘Metal cutting theory and practice’ (CRC Press, Boca Raton, USA, 2016).
[16] Sela, A., Ortiz-de-Zarate, G., Arrieta, I., et al.: ‘A mechanistic model to predict cutting force on orthogonal machining of 6061 aluminium alloy using GA for industrial applications’, Int. J. Mach. Machinability Mater., 2019, 21, (1–2), pp. 43–65.
[17] Abidin, N.W.Z., Ab Rashid, M.F.F., Mohamed, N.M.Z.N.: ‘A review of multi-hole drilling path optimization using soft computing approaches’, Arch. Comput. Methods Eng., 2019, 26, (1), pp. 107–118.
[18] Chandrasekaran, M., Muralidhar, M., Krishna, C.M., et al.: ‘Application of soft computing ‘predictive’ theories in machining, prediction and optimization: a literature review’, Int. J. Adv. Manuf. Technol., 2010, 46, (5–8), pp. 445–464.
[19] Rangwala, S.S., Dornfeld, D.A.: ‘Learning and optimization of machining operations using computing abilities of neural networks’, IEEE Trans. Syst. Man Cybern., 1989, 19, (2), pp. 299–314.
[20] Aozouri, R., Guillot, M.: ‘On-line prediction of surface finish and dimensional deviation: integration of MQL-assisted turning of Inconel 825 superalloy using GA for industrial applications’, Int. J. Mach. Machinability Mater., 2019, 21, (1–2), pp. 43–65.
[21] Simões, N.P., Cunha, A.C., Monteiro, H.F.: ‘A neural network based approach to prediction and control of cutting forces in turning processes’, Int. J. Adv. Manuf. Technol., 2019, 101, (9–12), pp. 3073–3080.
[22] Astakhov, V.P.: ‘Authentication of FEM in metal cutting’, in Paulo Davim, J (Eds.): ‘Finite element method for manufacturing processes: models, techniques and applications’, John Wiley & Sons, 2013, pp. 32–36.
[23] Vannatter, D., Bowles, J., Jay, W.P.: ‘Machine tool vibration and cutting force estimation’, CIRP Ann. Manuf. Technol., 1999, 48, (2), pp. 578–626.
[24] Tamang, S.K., Chandrasekaran, M., Palanikumar, K., et al.: ‘Machining performances’, Arch. Comput. Methods Eng., 2019, 26, (1), pp. 107–118.
[25] Chandrasekaran, M., Muralidhar, M., Krishna, C.M., et al.: ‘Application of soft computing ‘predictive’ theories in machining, prediction and optimization: a literature review’, Int. J. Adv. Manuf. Technol., 2010, 46, (5–8), pp. 445–464.
[26] Rangwala, S.S., Dornfeld, D.A.: ‘Learning and optimization of machining operations using computing abilities of neural networks’, IEEE Trans. Syst. Man Cybern., 1989, 19, (2), pp. 299–314.
[27] Aozouri, R., Guillot, M.: ‘On-line prediction of surface finish and dimensional deviation: integration of MQL-assisted turning of Inconel 825 superalloy using soft computing approaches’, Arch. Comput. Methods Eng., 2019, 26, (1), pp. 107–118.
[28] Rangwala, S.S., Dornfeld, D.A.: ‘Learning and optimization of machining operations using computing abilities of neural networks’, IEEE Trans. Syst. Man Cybern., 1989, 19, (2), pp. 299–314.
[29] Aozouri, R., Guillot, M.: ‘On-line prediction of surface finish and dimensional deviation: integration of MQL-assisted turning of Inconel 825 superalloy using soft computing approaches’, Arch. Comput. Methods Eng., 2019, 26, (1), pp. 107–118.
[30] Rangwala, S.S., Dornfeld, D.A.: ‘Learning and optimization of machining operations using computing abilities of neural networks’, IEEE Trans. Syst. Man Cybern., 1989, 19, (2), pp. 299–314.
[31] Aozouri, R., Guillot, M.: ‘On-line prediction of surface finish and dimensional deviation: integration of MQL-assisted turning of Inconel 825 superalloy using soft computing approaches’, Arch. Comput. Methods Eng., 2019, 26, (1), pp. 107–118.
[32] Rangwala, S.S., Dornfeld, D.A.: ‘Learning and optimization of machining operations using computing abilities of neural networks’, IEEE Trans. Syst. Man Cybern., 1989, 19, (2), pp. 299–314.