**Data processing and Manipulation:**

Data processing in machine learning is a very crucial step where we mould and manipulate data before feeding it to our model.

Clean and balanced data is important to make unbiased and correct prediction, there are multiple steps to be taken in order to train the model in right way, these may vary in importance with respect to specific models and their behaviour.

**Data Collection:** The dataset is divided into two major parts as ‘label\_data’ and ‘sensor\_data’ we are using and focusing on the later one;

In sensor\_data we have direct sensor readings of accelerometer, Gyroscope and Euler’s angle which are distributed in several small excel sheets labelled in a particular format as “SxxTyyRzz”; where S stands for **subject** and xx is the number denoting a subject (6 to 38) , T stands for **Task** and yy is the task id (1 to 36), R stands for **Retries/attempts** and zz denotes the no. of try/attempt (1 to 5)

Each data file has 11 columns as – Timestamp, Frame count, Acc and Gyroscopic measurement in all 3 dimensions and Euler’s angle in 3 dimensions

We have a very huge dataset with around 60 lakh entries.

**Data Sorting :** the dataset doesn’t explicitly showcases which one is data for falling motion and which one is non-falling motion, instead the first half of the dataset with task id 1 to 19 and 35, 36 are task where falling is not included and others are of course the opposite

| **Task ID** | **Description** |
| --- | --- |
| 01 | Stand for 30 seconds |
| 02 | Stand, slowly bend the back with or without bending at knees |
| 03 | Pick up an object from the floor |
| 04 | Gently jump (try to reach an object) |
| 05 | sit to the ground, wait a moment, and get up with normal speed |
| 06 | Walk normally with turn (4m) |
| 07 | Walk quickly with turn (4m) |
| 08 | Jog normally with turn (4m) |
| 09 | Jog quickly with turn (4m) |
| 10 | Stumble while walking |
| 11 | Sit on a chair for 30 seconds |
| 12 | Sit on the sofa (back is inclined to the support) for 30 seconds |
| 13 | Sit down to a chair normally, and get up from a chair normally |
| 14 | Sit down to a chair quickly, and get up from a chair quickly |
| 15 | Sit a moment, trying to get up, and collapse into a chair |
| 16 | Stand, sit on sofa (back is inclined to the support), and get up normally |
| 17 | Lie on the bed for 30 seconds |
| 18 | Sit a moment, lie down to the bed normally, and get up normally |
| 19 | Sit a moment, lie down to the bed quickly, and get up quickly |
| 35 | Walk upstairs and downstairs normally (5 steps) |
| 36 | Walk upstairs and downstairs quickly (5 steps) |
| 20 | Forward fall when trying to sit down |
| 21 | Backward fall when trying to sit down |
| 22 | lateral fall when trying to sit down |
| 23 | Forward fall when trying to get up |
| 24 | lateral fall when trying to get up |
| 25 | Forward fall while sitting, caused by fainting |
| 26 | lateral fall while sitting, caused by fainting |
| 27 | Backward fall while sitting, caused by fainting |
| 28 | Vertical(forward) fall while walking caused by fainting |
| 29 | Fall while walking, use of hands to dampen fall, caused by fainting |
| 30 | Forward fall while walking caused by a trip |
| 31 | Forward fall while jogging caused by a trip |
| 32 | Forward fall while walking caused by a slip |
| 33 | Lateral fall while walking caused by a slip |
| 34 | Backward fall while walking caused by a slip |

To work with them properly we sorted them explicitly in fall and no fall categories for every subject and added a column which will denote the fall condition (0 for no fall and 1 for fall)

We used pandas, OS and Numpy to make this possible

**Data cleaning** - First of all we will check for unwanted parameters which we will not need for our purpose,

In our case the Euler’s angles, Timestamp and Frame count are not so important and not needed ;

Timestamp and frame are details are related to the videos which show subjects performing tasks listed above.

Euler’s angle : The Euler angles are three angles introduced by Leonhard Euler to describe the orientation of a rigid body with respect to a fixed coordinate system, these angels heavily depend upon the initial orientation, our conclusion was to train with and without these params and check for results,

We found out it is better to drop these columns for our model.

In the remaining dataset of useful parameters we will search for any missing or faulty entries in the dataset and also for any kind of inconsistencies such as different data-types and impossible values;

Fortunately, our dataset didn’t suffer with any of the above problems.

**Data Balancing :** The dataset needed for prediction models need to be balanced with respect to target values, in our case it means that the data available for Fall and NO-Fall actions must be same in quantity .

Not having balanced data may even lead to high accuracy in some cases and pose problems in real life events and stay undetected if not dealt with beforehand because it can be considered as a logical error for our model and result in bias.

In our case we have minority of Fall subset with 1725407 entries verses the No-fall subset with 2269693 entries.

We have used Synthetic Minority Over-sampling Technique (SMOTE) to balance our data. This technique will take the data which is in minority and create new training examples to match the size of our bigger set.

**Method selection and Model Training:**

Selecting a suitable training method for our model will involve various factors such as the nature of data, the complexity, the available computational resources, and the characteristics we want model to have.

Considering we only want binary output from the model even basic ML Algorithms like Linear regression and better ML Algo. Like SVM (Support Vector Machine) might work as well, along with powerful DL Algo. Like Neural Networks and Random forest.

It should be taken in consideration that having such a big dataset may also pose challenges of high bias and high variance.

Neural Networks with enough layers and Random forests with enough depth have special tendency to decrease bias and variance, and neural networks often work best with heavy amount of data.

We will train our model on all these algorithms and choose the one with best results.

**Training:** After training our model and feeding our data on these methods, we have:

| Algorithm | Accuracy | Remark |
| --- | --- | --- |
| Linear Regression | 78% | Better alternate available |
| Support Vector Machine | 74% | Needs more computational resource |
| Neural  Networks  (w/o Euler’s angle) | 81% | By far best  (can be improved) |
| Neural  Networks  (With Euler’s angle) | Below 70% | Too many input param changing fast affected NN Badly |

So, it seems that for our dataset and vision of model Neural Networks are suited best.