Encrypted Classification Using Secure K-Nearest Neighbor Computation

B Pradeep Kumar Reddy

Advanced Technology Development Centre,
IIT Kharagpur

Dr. Ayantika Chatterjee

Advanced Technology Development Centre,
IIT Kharagpur

Outline

Cloud computing

- Software as a Service (SaaS)
- Platform as a Service (PaaS)
- Infrastructure as a Service (IaaS)
- Machine Learning as a Service (MLaaS)

Characteristics of Cloud Computing

- On demand services
- N/w access
- Shared resources
- Scalability

What's the security bottleneck of this shared platform?

Famous Cloud/ Server Data Hacks

Organization	Impact
Yahoo(2013-14)	3 billion user accounts were compramised with names, Date of births, email addresses and passwords
Marriott International(2014-18)	attackers were able to take some combination of contact info, passport number and other personal information of 500 million customers from internal server
eBay (May 2014)	145 million users compromised when attackers got access to company network for 229 days using the credentials of three corporate employees
Equifax 2017	Personal information and Credit card data hacking of 143 million consumers of this one of the largest credit bureaus in the U.S
Hearland Payment Systems(2008)	134 million credit cards exposed through SQL injection, company paid out an estimated 145 million in compensation
Uber (2016)	Personal information of 57 million Uber users and 600,000 drivers exposed along with driver license num-bers

Famous Cloud/ Server Data Hacks

Organization	Impact
Home Depot (2014)	Theft of credit/debit card information of 56 million customers, company agreed pay at least 19:5 million to compensate US consumers
Micro Soft(2010)	Experienced breach due to configuration issue within Business Productivity Online Suite, allowed non-authorized users of the cloud service to access employee contact info
Drop Box(2014)	Hackers tapped more than 68 million user accounts and passwords, disclosed after four years
LinkedIn(2012)	6 million user passwords were stolen then published on a Russian forum
LinkedIn(2016)	hackers stole and posted for sale on the dark web an estimated 167 million LiknedIn email addresses and pass-words
Apple Icloud(2014)	high-profile cloud security breach, the iCloud service for personal storage of celebrities had been compromised, private photos leaked online

Cryptographic algorithms

- Traditional Algorithms
- RSA
- Elliptic Curve Cryptography (ECC)
- El Gamal
- Digital Signature Algorithm (DSA)
- Advanced Encryption Standard (AES)

Advantage:

Data is stored in Encrypted form

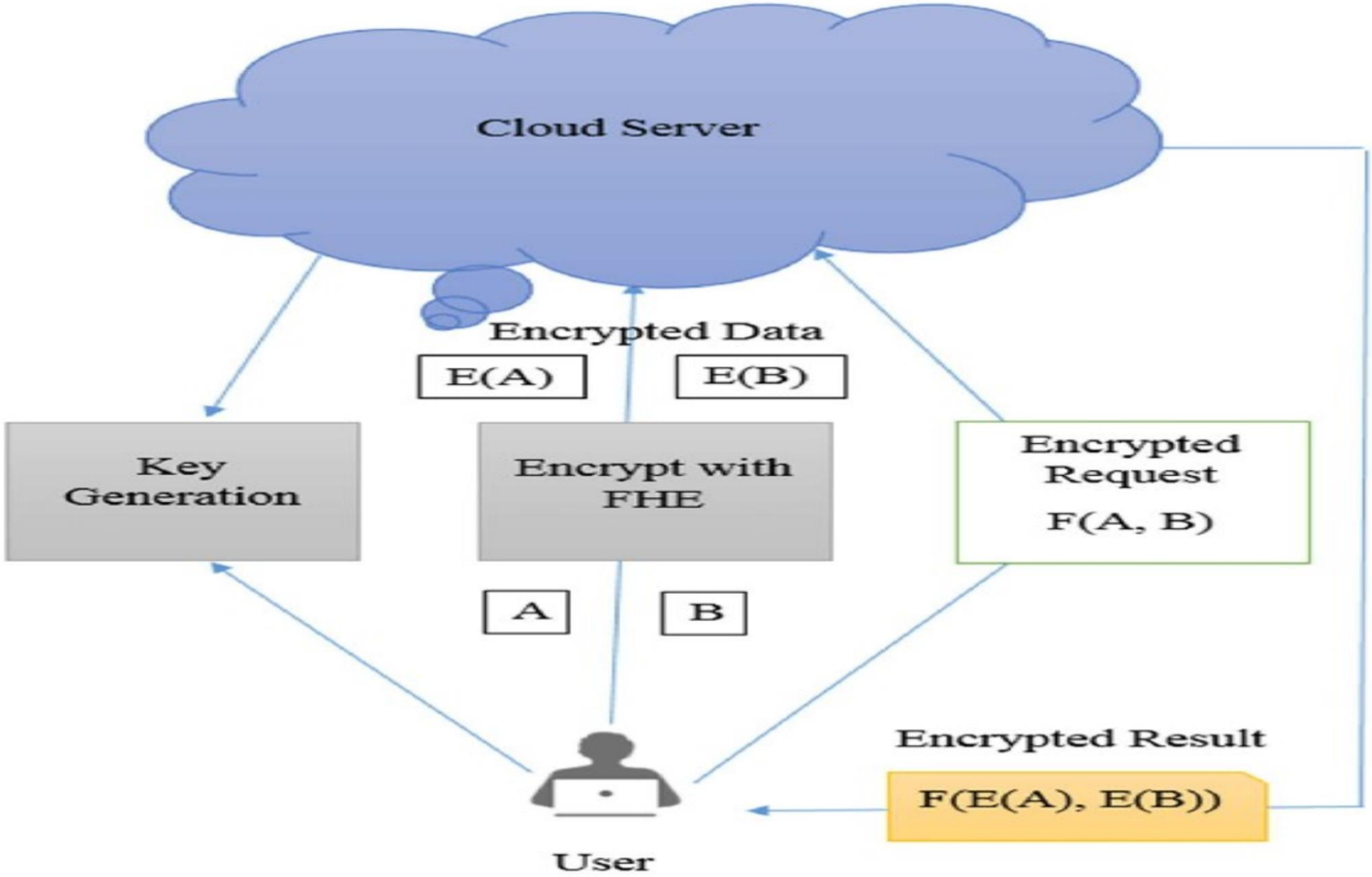
Disadvantage:

We can't Perform operations(like Addition , Subtraction…) on Encrypted data using above algorithms

Solution

- Due to public access to the information in clouds, security is a major challenge.
- Encryption of stored data can preserve confidentiality, but reduces speed in case of performing any operation on encrypted data.
- **Is it possible to delegate processing of data without giving access to it?**
- Only solution : **Homomorphic encryption on stored data**.

Homomorphic Encryption



A Homomorphic encryption scheme

- Shared secret key: p
- To encrypt a bit: m
- Choose at random large q, small r
- Output $c = pq + 2r + m$
- **(Ciphertext is close to a multiple of** p
- m **= LSB of distance to nearest multiple of** p **)**
- To decrypt : $m = c \bmod p$

How homomorphic is the scheme?

- C1=q1p+2r1+m1, C2=q2p+2r2+m2
- C1+C2=(q1+q1)p+2(r1+r2)+(m1+m2)

distance to nearest multiple of p

(C1+C2)mod p= 2(r1+r2)+(m1+m2)

error term

- C1*C2=(C1q2+C2q1-q1q2)p+2(2r1r2+r1m2+m1r2)+m1m2
- (C1*C2)mod p= 2(2r1r2+r1m2+m1r2)+m1m2

error term

Homomorphic property retains till the error-term is within a certain limit.

→ Noise doubles on addition, squares on multiplication.

Homomorphic encryption schemes

- **Goldwasser-Micali encryption (1982)**: additive homomorphism.
- **ElGamal encryption (1984)**: Multiplicative homomorphism.
- **Paillier encryption (1999)**: Multiplicative homomorphism.
- All the above schemes are partially homomorphic.
- Gentry first proposed a solution as **Fully homomorphic encryption** scheme (2009) [1].

Our Motivation

“it must be emphasized that homomorphy is a theoretical achievement that merely lets us arithmetically add and multiply plaintexts encapsulated inside a ciphertext. In theory, this allows the execution of any algorithm complex manipulations like text replacements or similar, but putting this to practice requires the design (compilation) of a specific circuit representation for the algorithm at hand. This may be a nontrivial task.” [11]

Machine learning

- **Supervised learning**
  - -Classification
  - -Regression
- **Unsupervised learning**
  - -Clustering

How to realize the encrypted counterpart of these learning algorithms?

Why KNN?

- KNN is simple, easy to understand, and implement. To classify the new

datum KNN reads through the whole dataset to search out K nearest neighbours.

- No assumptions need to make to implement kNN. Parametric models have

a lot of assumptions.

- KNN does not explicitly build any model, merely tags the new data entry
- based learning from historical data.
- KNN complies with multi-class with none additional efforts.

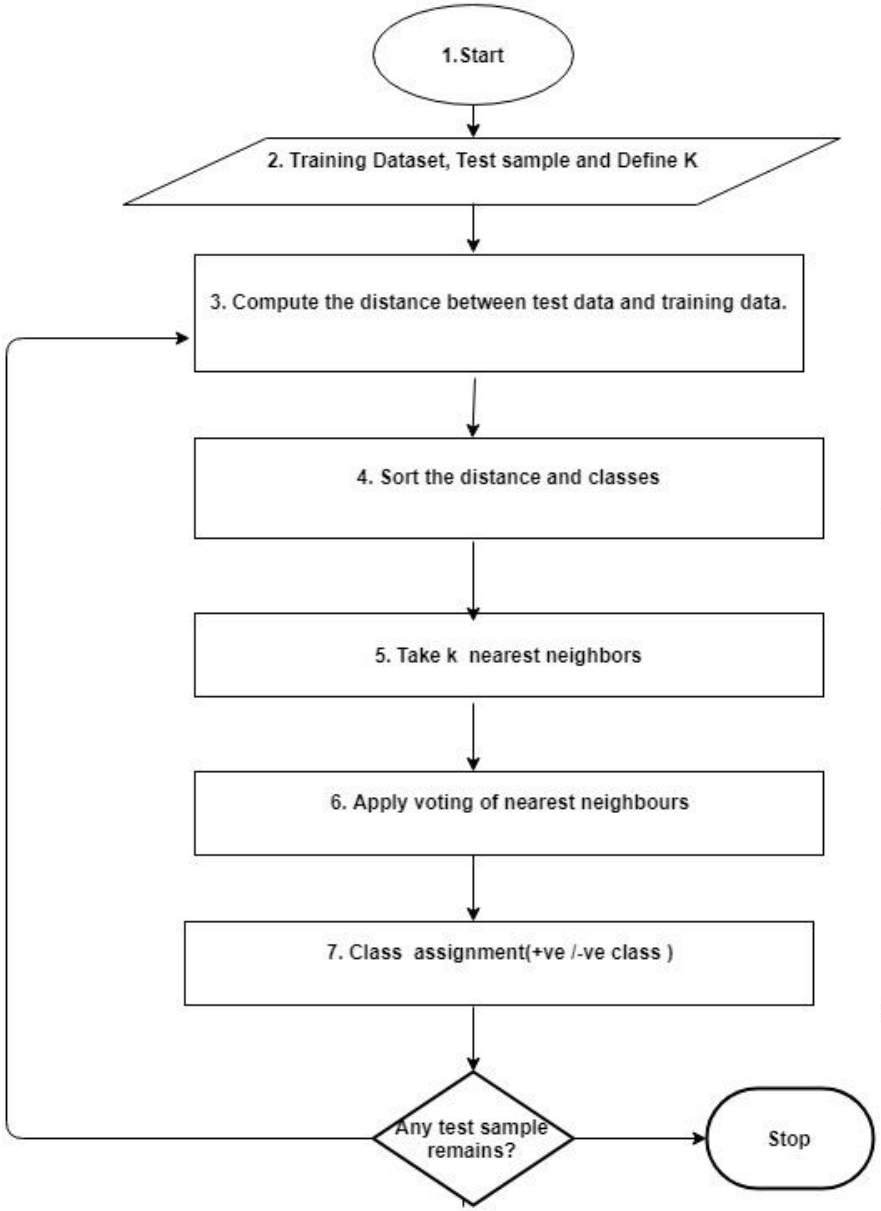
Previous works

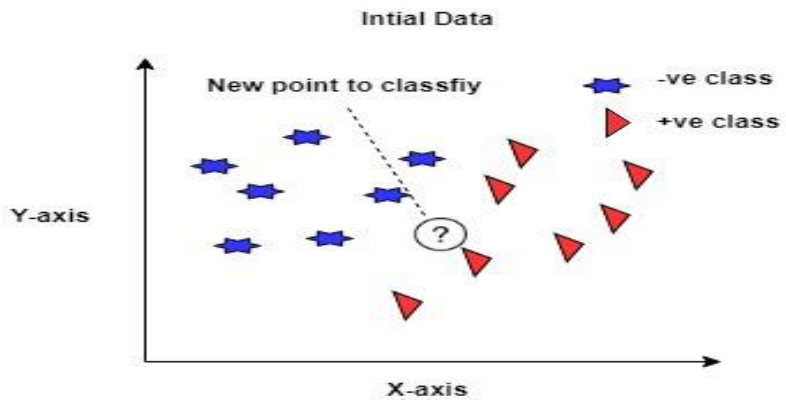
- particular KNN Queries implementation with Location

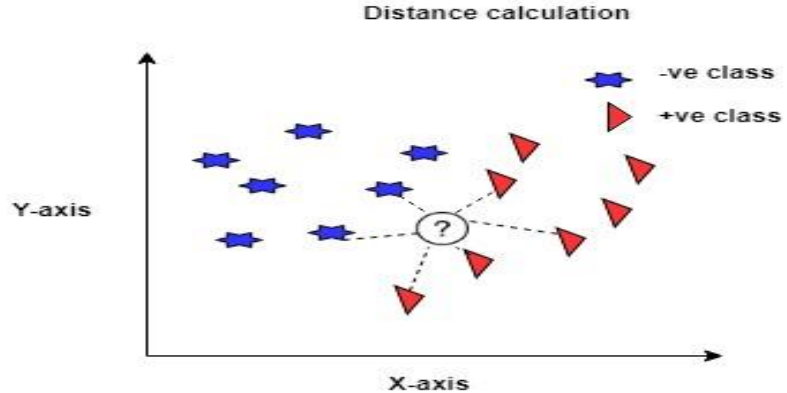
privacy useful for mobile communication [3]

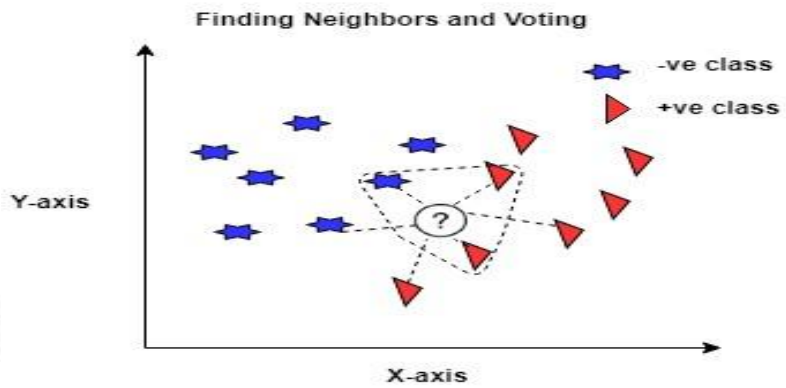
- Or secure KNN query processing based on mutable order preserving encoding (mOPE) with inherent limitations [4]
- Work in [5] performs only the secure KNN search over encrypted data using searchable encryption scheme.
- In [7], authors proposed secure KNN design with restricted homomorphic property.
- Recent work in [6] demonstrates secure KNN computing in two-party federated cloud setting requires one round of communication between the two.

KNN Algorithm









Key Operations

For the algorithm with input as:

- n point training data set P ={p1, p2, p3, . . . pn} with m dimensional features and associated class labels
- number of nearest neighbours integer value K and a test data T

- Distance computation between the points p_i and T_i.
- Distance sorting
- Class label assignment

Distance computation

- Euclidean distance
- $d(x, y) = \sqrt{\sum_{i=1}^{n}(attr_i(x) - attr_i(y))}\text{^}2$

  x , y are n-dimensional vectors
- Minkowski distance

  $d(x, y) = \ (\sqrt[p]{\sum_{i=1}^{n} abs(attr_i(x) - attri(y))}\ \text{^}p$
- Manhattan distance

  $\mathrm{d}(x, y) = \ \sum_{i=1}^{n} abs(attr_i(x) - attri(y)$

Distance computation: FH subtraction

- Homomorphic subtraction with a^j and b^j (encryptions of a and b respectively) is defined as:

$$a^j - b^j = a^j + Encrypt(2^j s\ complement\ of\ b)$$

- The 2's complement of b in the encrypted domain is obtained as follows:
- $Encrypt((2^j s\ complement\ of\ b), pk) = b^j \oplus Encrypt(11\cdots1, pk) \oplus Encrypt(1, pk)$

Absolute distance computation

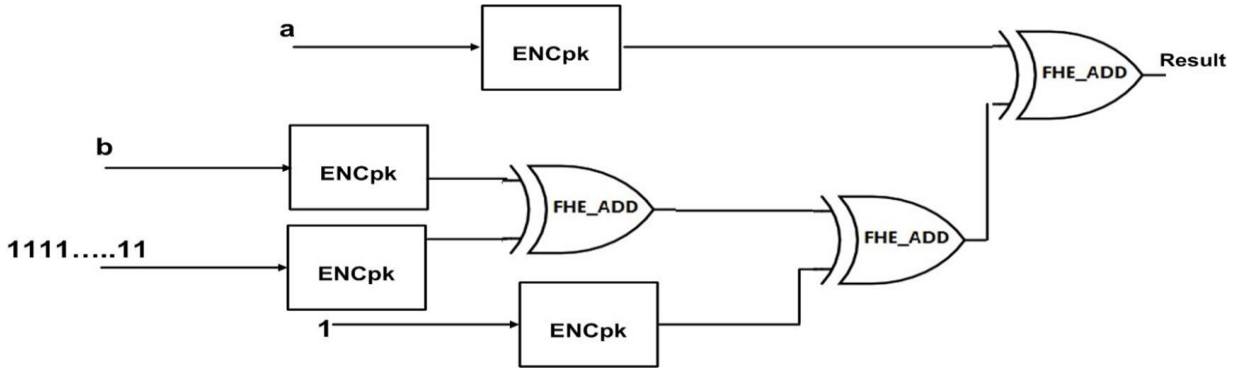


Figure : Fully Homomorphic Subtraction

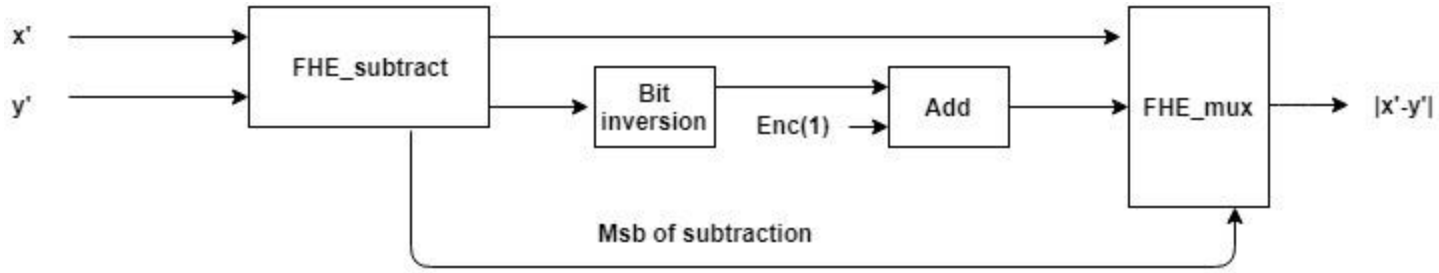
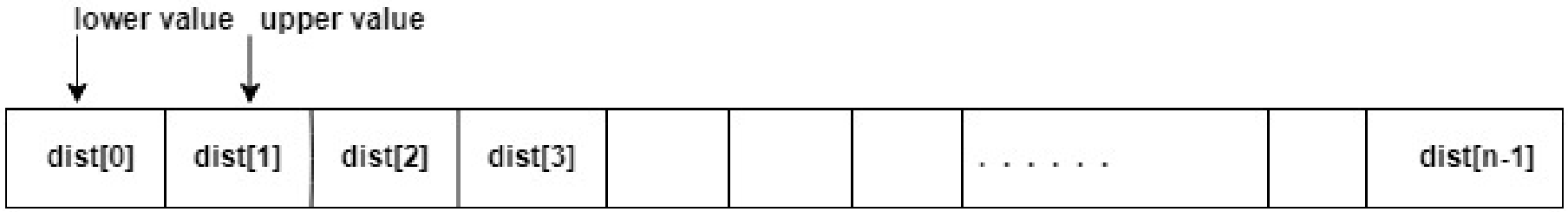


Figure : Fully Homomorphic absolute distance computation

Distance Sorting



Sorting of computed distances requires implementation of encrypted sorting with dist'[i] values.

Sorting Choices

- Bubble Sort over Partition based sort (Quick sort and Merge sort)

Why comparison based sorting is secured?[2]

- - To implement comparison based sort, it is never revealed whether the swapping is actually taking place. Only greater and smaller elements maintain their proper positions.
- In partition based sort, it is essential to know whether an element is really greater or smaller compared to the pivot (partitioning condition).

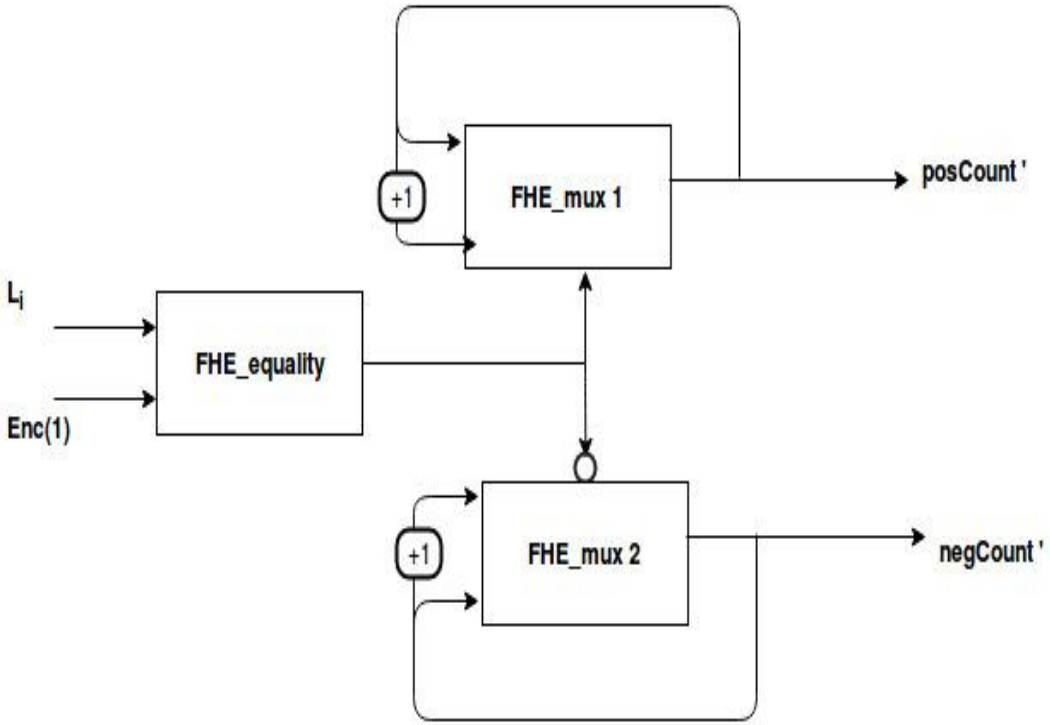
KNN Voting

- From the final sorted distance array, K nearest neighbours are selected
- final assignment of test input's class label will be done by **KNN Voting**

- Training phase class labels are assigned either as positive (+) class

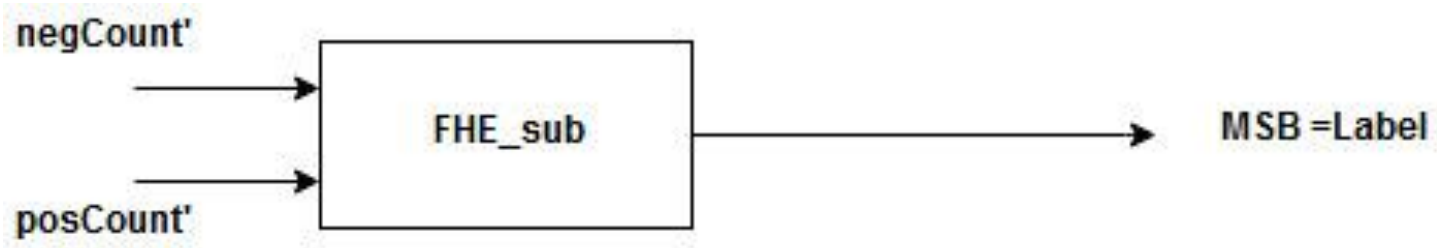
or negative class marked as label (L_i) and encrypted as Enc(1) or Enc(0).

Encrypted Prediction



- Equality comparison of nearest neighbours class labels (L_i) with Enc(1) or Enc(0).
- Summing up the the equality check results to get the total positive (posCount') and negative counts (negCount').

Class Label Assignment



If negCount$_i$ > posCount$_i$:
MSB turns to be Enc(0) and that predicts the negative class, else otherwise.

Results

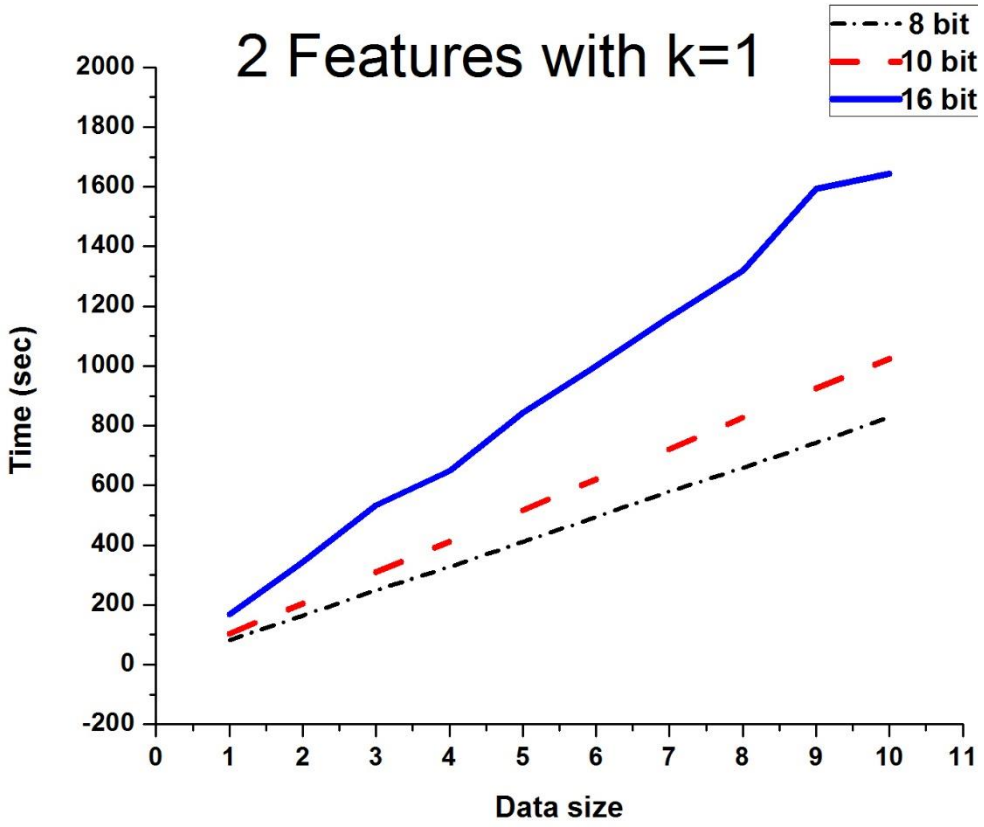
Implementation

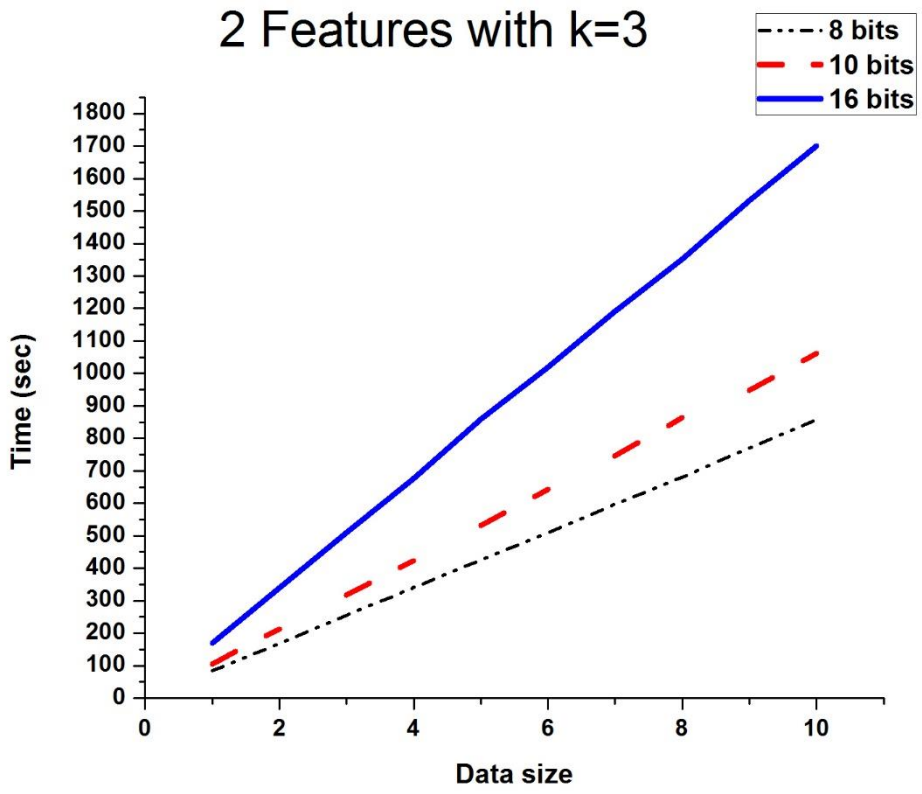
- Iris Dataset from UCI machine learning repository
- TFHE module
- Ubuntu 64-bit machine with i7770 with 3.6 GHZ processor implementation without any parallel processing support so far.
- Iris Dataset Desription

1. sepal length
2. sepal width
3. petal length
4. petal width
5. class:

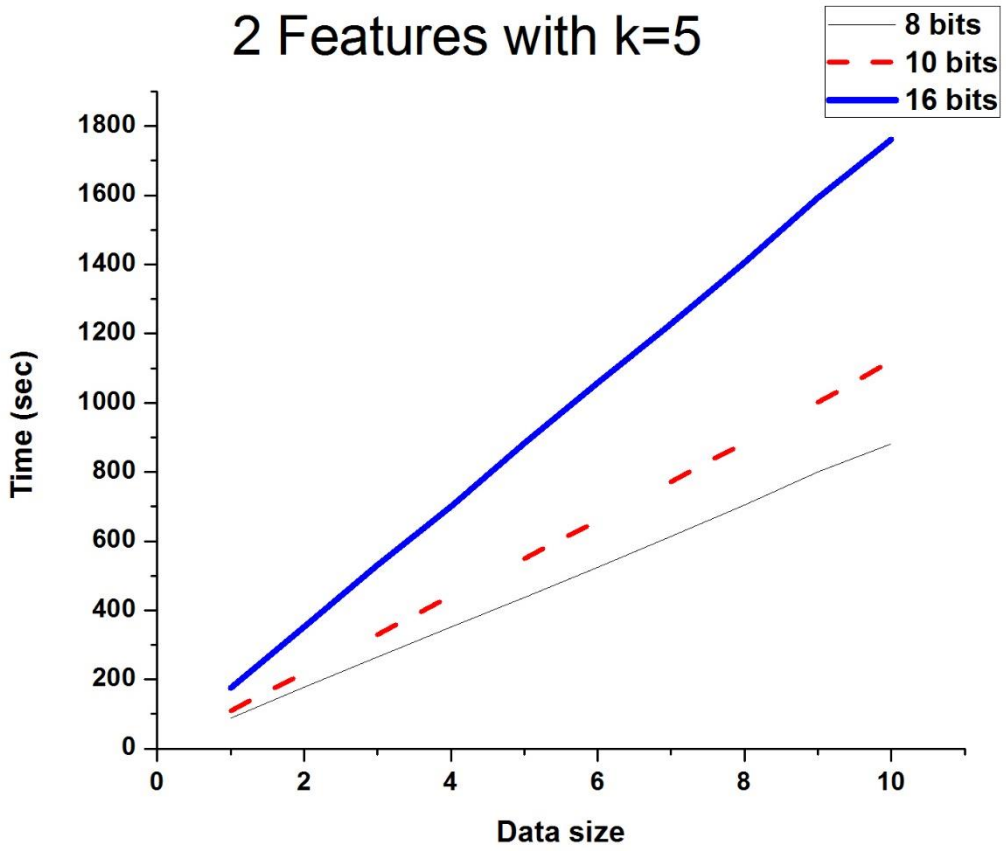
-- Iris Setosa
-- Iris Versicolour
-- Iris Virginica

Data size vs Time

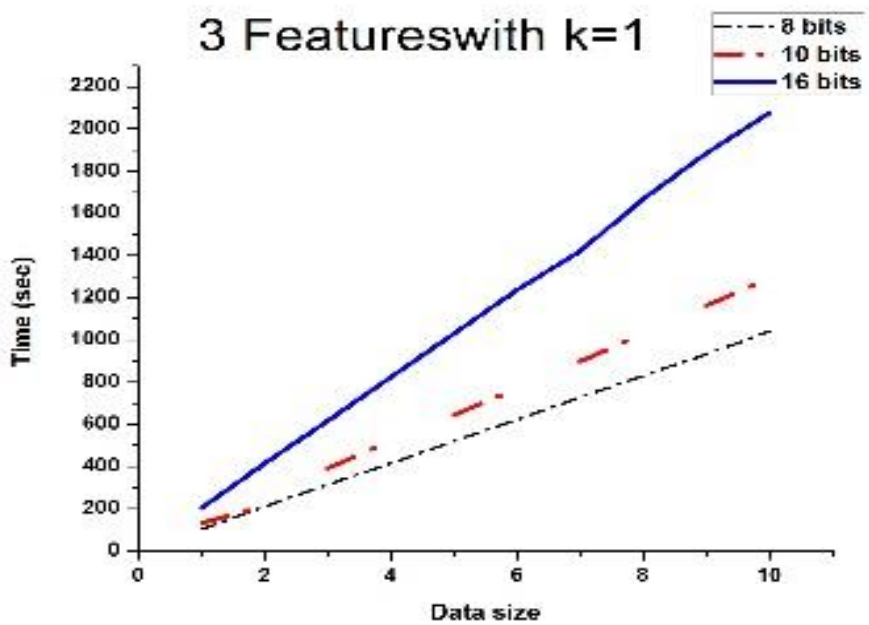


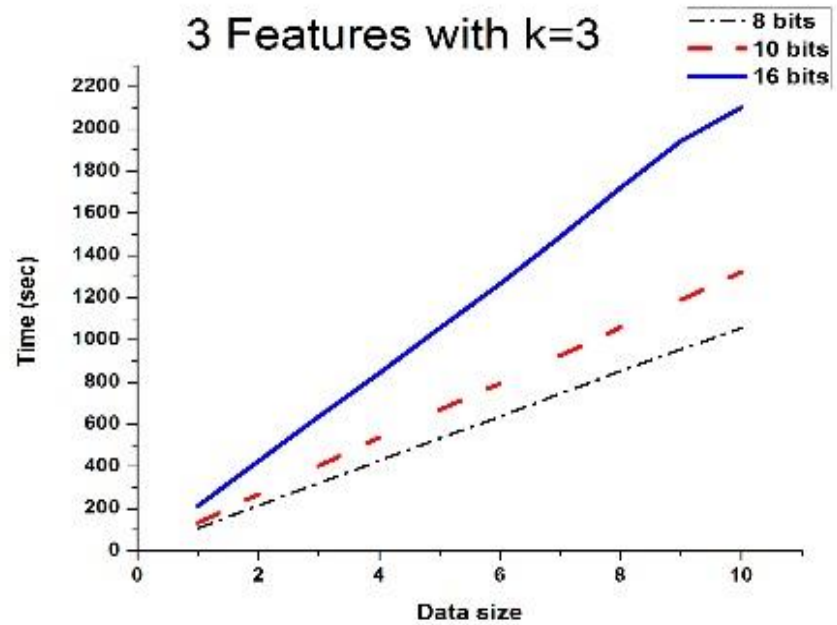


Data size vs Time

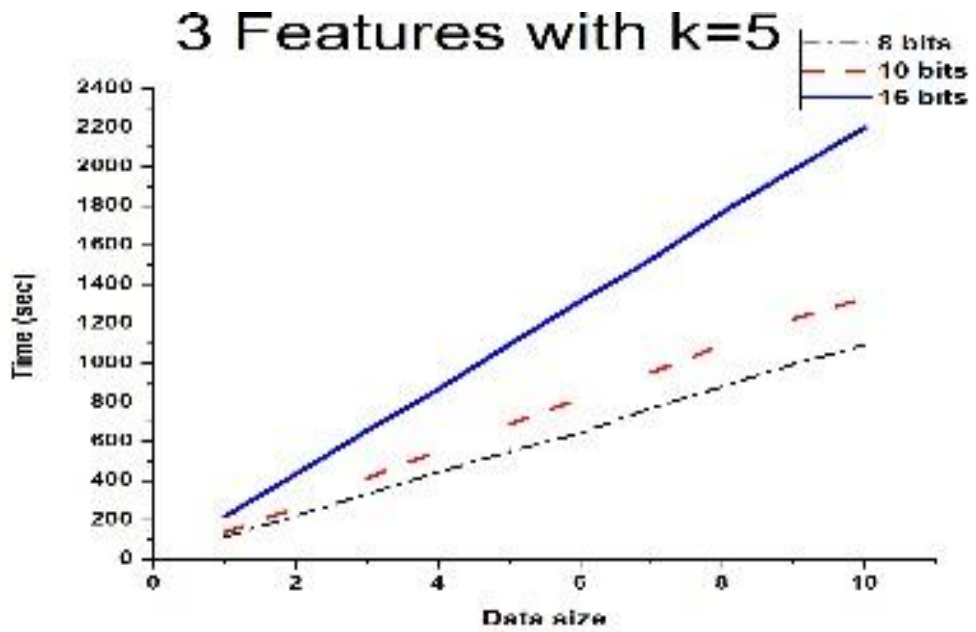


Data size vs Time

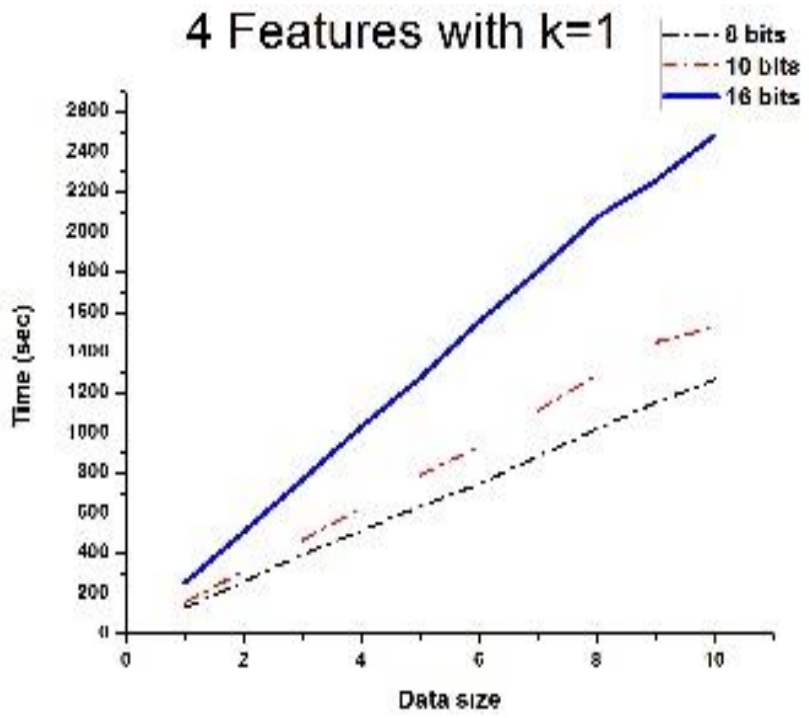


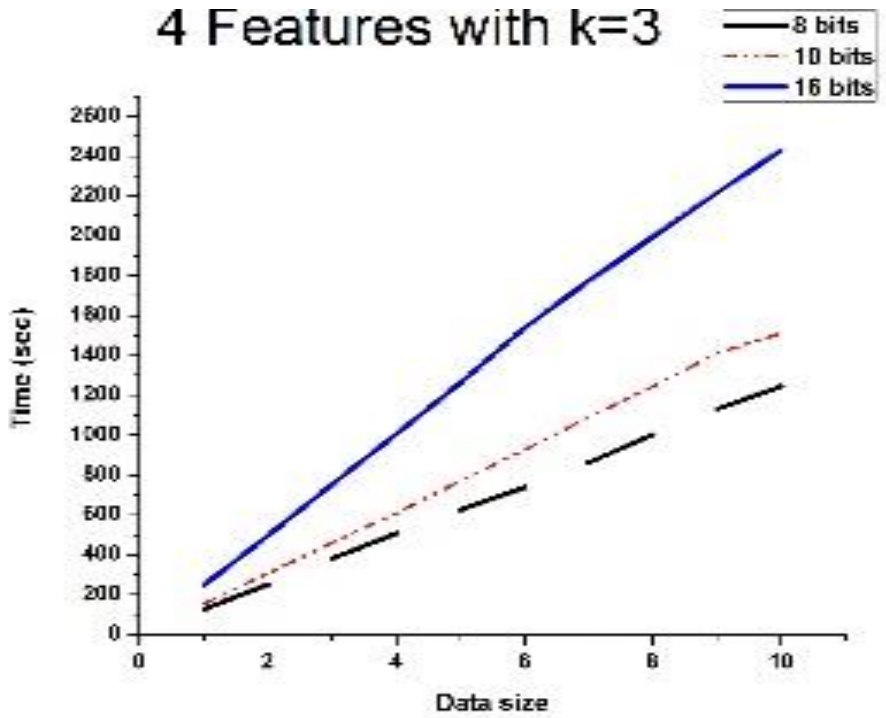


Data size vs Time

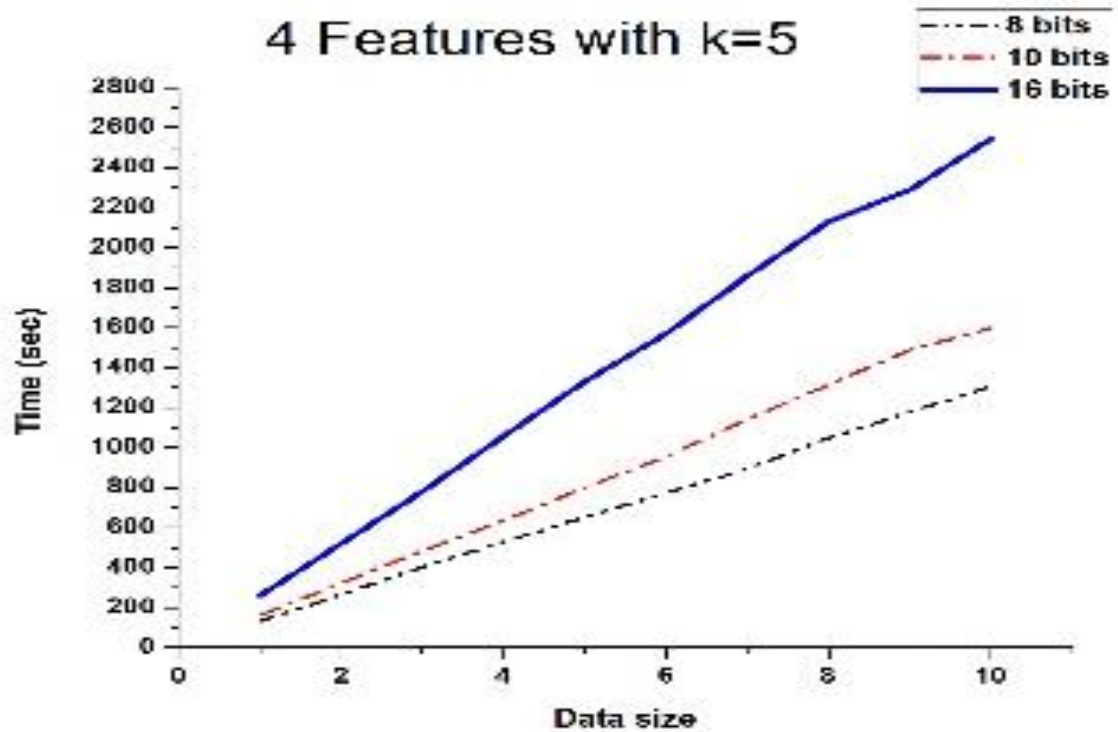


Data size vs Time

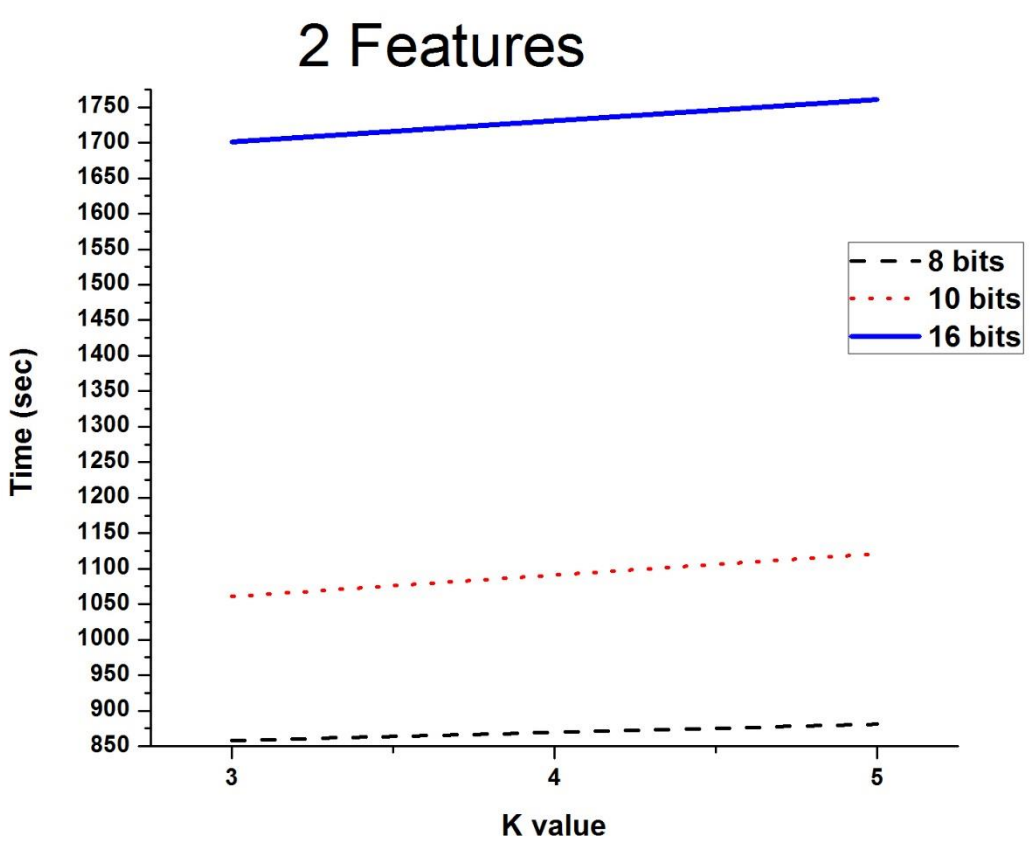


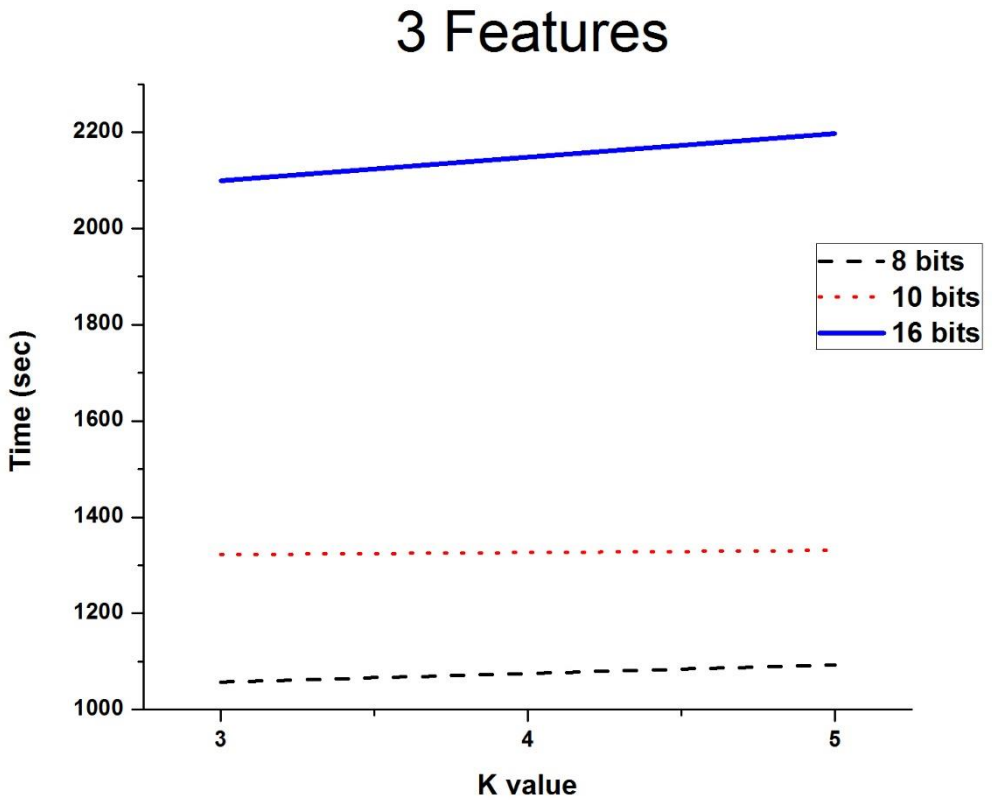


Data size vs Time

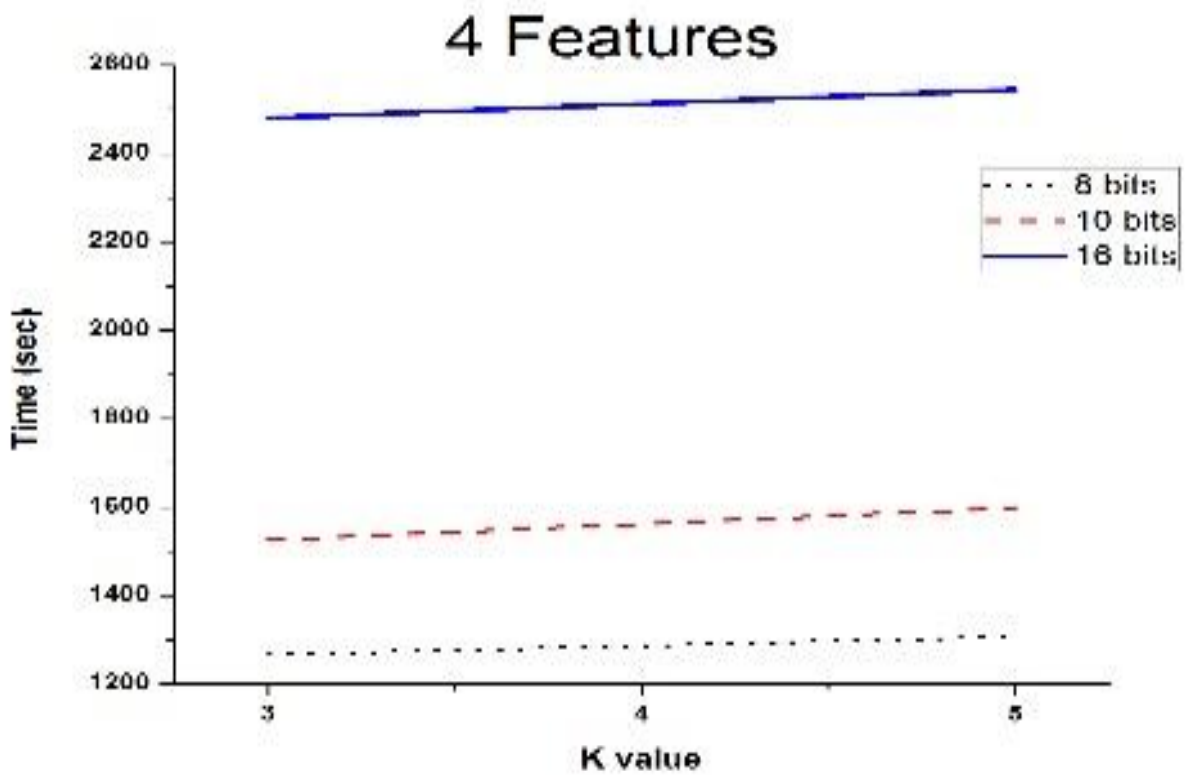


K vs Time

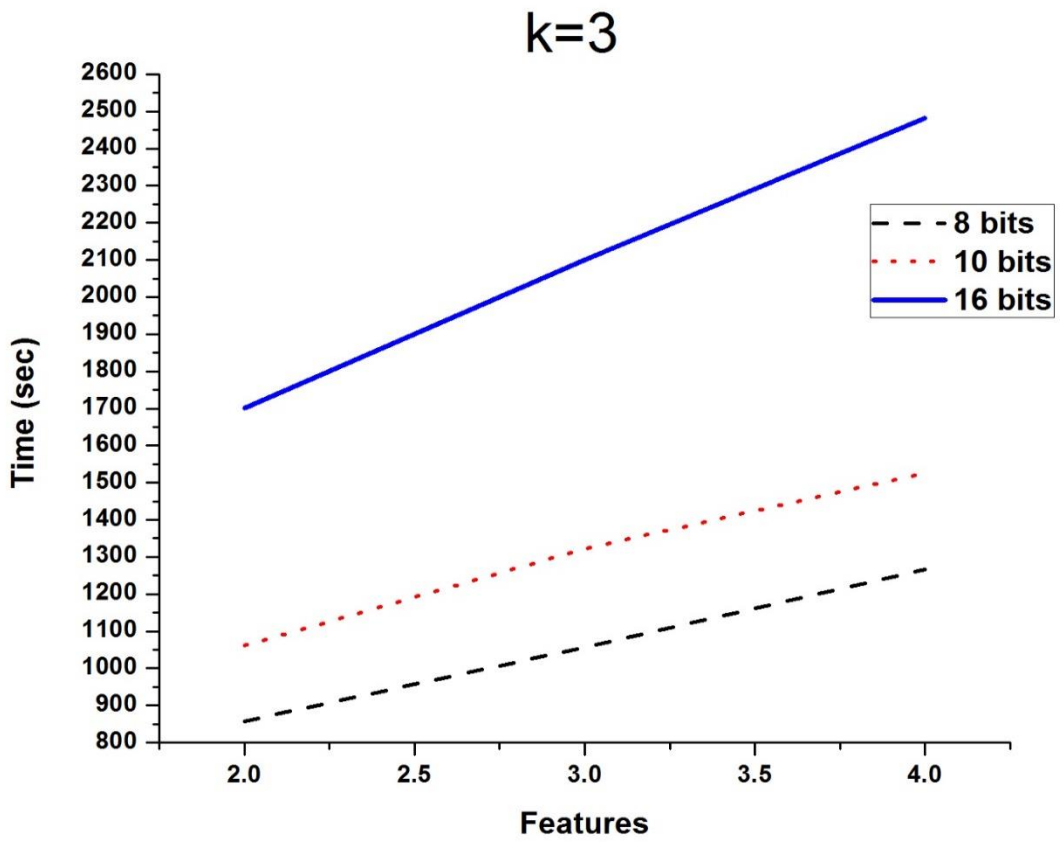

2 Features
Time (sec)
1750
1700
1650
1600
1550
1500
1450
1400
1350
1300
1250
1200
1150
1100
1050
1000
950
900
850
3
4
5
K value
8 bits
10 bits
16 bits

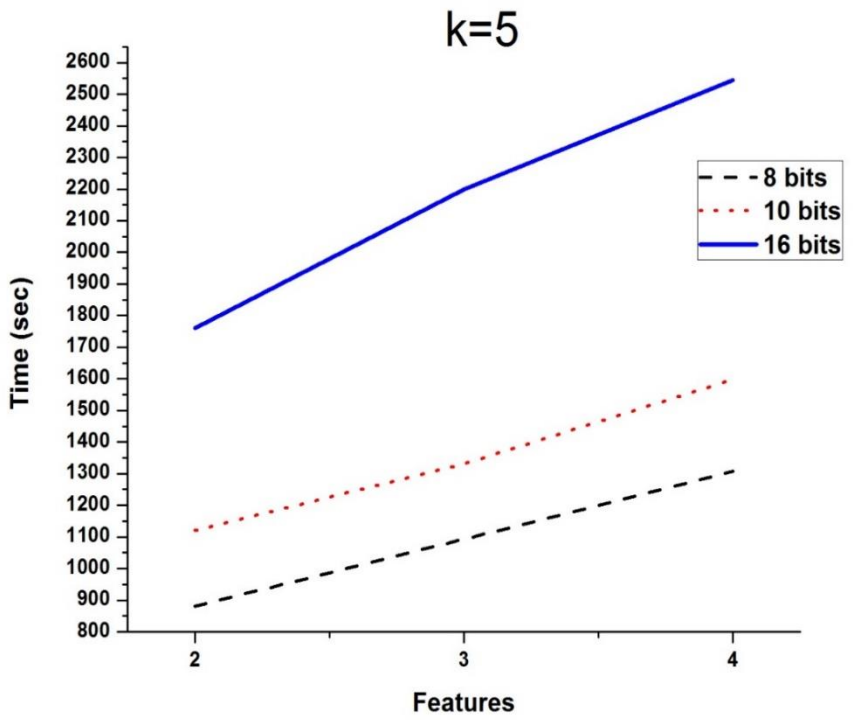

3 Features
Time (sec)
2200
2000
1800
1600
1400
1200
1000
3
4
5
K value
8 bits
10 bits
16 bits

K vs Time


4 Features
2600
2400
2200
2000
1800
1600
1400
1200
Time (sec)
3
4
5
K value
8 bits
10 bits
16 bits

Features vs Time





Conclusion

- Privacy preserving machine learning can be effective for critical medical ,financial data and long-term decision making applications

- Future scope
- Performance improvement supporting suitable parallel processing and GPU or FPGA acceleration of the underlying library.

- Suitable encrypted operator design for other classification and regression algorithms.

References

- 1. Van Dijk, M., Gentry, C., Halevi, S., & Vaikuntanathan, V. (2010, May). Fully homomorphic encryption over the integers. In *Annual International Conference on the Theory and Applications of Cryptographic Techniques* (pp. 24-43). Springer, Berlin, Heidelberg..
- 2. Chatterjee, Ayantika and Indranil Sengupta:Searching and Sorting of Fully Homomorphic Encrypted Data on Cloud. IACR Cryptology ePrint Archive 2015 (2015): 981.
- 3. Yi, Xun, Russell Paulet, Elisa Bertino and Vijay Varadharajan: Practical k nearest neighbor queries with location privacy. In: IEEE 30th International Conference on Data Engineering (2014): 640-651..

- 4. Elmehdwi, Yousef, Bharath K. Samanthula and Wei Jiang: Secure k-nearest neighbor query over encrypted data in outsourced environments. In: IEEE 30th International Conference on Data Engineering (2014): 664-675.
- 5. Wang, Boyang, Yantian Hou and Ming Li: Practical and secure nearest neighbor search on encrypted large-scale data. In: IEEE INFOCOM 2016 - The 35th Annual IEEE International Conference on Computer Communications (2016): 1-9.
- 6.Kesarwani,Manish,AksharKaul,PrasadNaldurg,SikharPatranabis,Gagandeep Singh, Sameep Mehta and Debdeep Mukhopadhyay:Efficient Secure k-Nearest Neighbours over Encrypted Data. EDBT (2018),DOI:10.5441/002/edbt.2018.67.
- 7. Zhu, Youwen, Zhiqiu Huang and Tsuyoshi Takagi: Secure and controllable kNN query over encrypted cloud data with key confidentiality: Parallel Distrib. Comput. 89 (2016): 1-12

- 8. Chen, Hao, Ran Gilad-Bachrach, Kyoohyung Han, Zhicong Huang, Amir Jalali, Kim Laine and Kristin E. Lauter: Logistic regression over encrypted data from fully homomorphic encryption. BMC Medical Genomics (2018).
- 9. Hu, Shengshan, Qian Wang, Jingjun Wang, Sherman S. M. Chow and Qin Zou: Securing Fast Learning! Ridge Regression over Encrypted Big Data. 2016 IEEE Trustcom/BigDataSE/ISPA (2016): 19-26.
- 10. Laur, Sven and Lipmaa, Helger and Mielikinen, Taneli: Cryptographically private support vector machines. In: 12th ACM SIGKDD international conference on Knowledge discovery and data mining,2016.
- 11. Stefan Rass ,Danieal Slamanig: Cryptography for Security and Privacy in Cloud Computing

THANK YOU